Title
Time-varying networks: Measurement, Modeling, and Computation

Author
Yu, Yue

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Time-varying networks: Measurement, Modeling, and Computation

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Yue Yu

Dissertation Committee:
Professor Carter T. Butts, Chair
Professor Katherine Faust
Professor Athina Markopoulou

2019

DEDICATION

This dissertation is wholeheartedly dedicated to my parents, Zhongqin Yu and Jing Xiong, for their unconditional love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first and foremost like to thank my Ph.D. advisor, Professor Carter Butts, for all of his advice, encouragement, and wisdom. He has made me, and many others, better scientists.

I would also like to thank the members of my dissertation committee, Professor Katherine Faust, Professor Athina Markopoulouand, and the two additional members of my advancement committee, Professor Charless Fowlkes, and Professor Michael Dillencourt for their helpful suggestions and advice.

I thank the members of the NCASD Lab, for the insightful conversations and collaboration.

This material is based on research supported by NIH 1R01HD068395-01, ARO award W911NF-14-1-0552, and NSF awards IIS-1526736 and DMS-1361425.

CURRICULUM VITAE

Yue Yu

Education

University of California, Irvine **2014 - 2019**
Ph.D. in Computer Science
Dissertation: "Time-varying networks: Measurement, Modeling and Computation"

University of California, Irvine **2011 - 2013**
M.S. in Computer Science, GPA 3.96/4.0

Hong Kong University of Science and Technology **2006 - 2010**
B.Eng. in Electrical Engineering (Information & Telecommunication)

Research Experience

University of California, Irvine. Irvine, CA, USA **2014 - present**
Ph.D. Researcher, Networks, Computation, and Social Dynamic Lab

Hong Kong University of Science and Technology, Hong Kong **2010 - 2011**
Researcher, The Human Language Technology Center

Hong Kong University of Science and Technology, Hong Kong **2019 - 2010**
Undergraduate Researcher

Teaching Experience

University of California, Irvine **2016, 2018**
Teaching Assistant, Introduction to Artificial Intelligence

Honors and Awards

6th Place, Kaggle Competition "Give Me Some Credit", Kaggle **2012**

1st Runner-up, IEEE Student Paper Contest (Region 10), *IEEE Region* 10 **2010**

- "Modified-MMSE Equalization for Noise and Interference Mitigation in SU-MIMO Systems"

2nd Runner-up, Best Final Year Project Awards, *Dept. of Electrical and Computer Engineering* **2010**

- Thesis: "Signal Processing for Interference Mitigation in MIMO Networks"

First Class Honors, Hong Kong University of Science and Technology **2010**

Dean's List, Hong Kong University of Science and Technology **2008 - 2010**

Provost's Honors, University of California, San Diego **2009**

Department Scholarship, Hong Kong University of Science and Technology **2006**

Referred Journal Publications

Retrospective Network Imputation from Life History Data: the Impact of Designs,
Sociological Methodology (2020).

Network-based Classification and Modeling of Amyloid Fibrils,
Journal of Physical Chemistry (ACS) (2019),
https://pubs.acs.org/doi/abs/10.1021/acs.jpcb.9b03494

Retweeting Risk Communication: the Role of Threat and Efficacy,
Risk Analysis (2018),
https://onlinelibrary.wiley.com/doi/full/10.1111/risa.13140

Conference Presentations

Local Graph Stability in Exponential Random Graph Models,
2st North American Social Networks Conference, 2018

Scalable Estimation for DNR Models of Sexual Contact Networks from Retrospective Life History Data,
37th International Sunbelt Network Conference, 2017

Concurrency Imputation from Egocentric Sexual History Data,
36th International Sunbelt Network Conference, 2016

ABSTRACT OF THE DISSERTATION

Time-varying networks: Measurement, Modeling, and Computation

By

Yue Yu

Doctor of Philosophy in Computer Science

University of California, Irvine, 2019

Professor Carter T. Butts, Chair

Time-varying networks and techniques developed to study them have been used to analyze dynamic systems in social, computational, biological, and other contexts. Significant progress has been made in this area in recent years, resulting from a combination of statistical advances and improved computational resources, giving rise to a range of new research questions. This thesis addresses problems related to three lines of inquiry involving dynamic networks: data collection designs; the conditions needed for structural stability of an evolving network; and the computational scalability of statistical models for network dynamics. The first contribution involves a commonly neglected problem concerning data collection protocols for dynamic network data: the impact of in-design missingness. A systematic formalization is offered for the widely used class of retrospective life history designs, and it is shown that design parameters have nontrivial effects on both the quantity of missingness and the impact of such missingness on network modeling and reconstruction. Using a simulation study, we also show how the consequences of design parameters for inference vary as a function of look-back time relative to the time of measurement. The second contribution of this thesis is related to a fundamental question of network dynamics: when or where are changes in a network most likely to occur? A novel approach is taken to this question, by exploring its complement – what factors *stabilize* a network (or subgraphs thereof) and make it resistant to change? For networks whose behavior can be parameterized in exponential

family form, a formal characterization of the graph-stabilizing region of the parameter space is shown to correspond to a convex polytope in the parameter space. A related construction can be used to find subgraphs that are or are not stable with respect to a given parameter vector, and to identify edge variables that are most vulnerable to perturbation. Finally, the third contribution of this thesis is to scalable parameter estimation for a class of temporal exponential family random graph models (TERGM) from sampled data. An algorithm is proposed that allows accurate approximation of maximum likelihood estimates for certain classes of TERGMs from egocentrically sampled retrospective life history data, without requiring simulation of the underlying network (a major bottleneck when the network size is large). Estimation time for this algorithm scales with the data size, and not with the size of the network, allowing it to be employed on very large populations.

Chapter 1

Introduction

1.1 Introduction

The last several decades have seen a growth of interest in network data and in strategies to analyze such data. For example, online social network (OSN) platforms such as Facebook and Twitter generate an enormous amount of social network data daily. This phenomenon motivates both empirical studies of network structure and the development of efficient measurement and modeling techniques to support the needs of OSN providers. While statistical analysis of social networks goes back to the 1930s [Moreno and Jennings, 1938], advances in computing and statistical theory have fueled a particular rise in stochastic models for networks with complex structure [Watts, 2004]. The importance of capturing nontrivial aspects of network structure has been motivated by studies of phenomena such as information transmission [e.g. Boorman, 1975, Dodds et al., 2003, Cowan and Jonard, 2004], systemic robustness [e.g. Krackhardt and Stern, 1988, Callaway et al., 2000, Klau and Weiskircher, 2005, Acemoglu et al., 2015], and biological assembly [e.g. Grazioli et al., 2019b], all of which can be significantly impacted by features such as clustering and community or subgroup

structure that are not readily reproduced by simple random graph models.

Researchers of the network analysis often encounter networks that are not static. The structures of the network may change over time, with either a nodal change (e.g., actors being in/out of the network, or properties associated with actors being altered), or an edgewise change (e.g., establishing a connection between actors, or breaking or changing existing connections). These rich dynamics reveal underlying mechanics that are fundamental to network analysis. For example, Morris and Kretzschmar [1997], Morris et al. [2009], show that the partnership concurrency is closely related to forward connectivity in time-varying networks. Broekel and Bednarz [2018] investigate the factors and their influence on the formation and dissolution of links, contributing to the analysis of the evolution of spatial (knowledge) networks.

Because of the complex nature of the time-varying networks, which involves both global and local structural properties, the dynamics of vertices and edges, and time-dependent components, many challenges arise from the studies of such networks. In this thesis, we aim to address three aspects of these challenges: measurement, modeling, and computation. In the rest of this chapter, we first introduce (in section 1.2) some fundamental concepts that are essential to the understanding of this thesis, then we will outline the succeeding chapters in section 1.3 and summarize our contributions.

1.2 Background

1.2.1 Network Data Representation

Network data are often described in the language of graph theory. A graph can be represented by an ordered pair $G = (V, E)$, where V is a set of vertices (also referred to as nodes),

and E is a set of edges which are pairs of distinct vertices, ordered or unordered. E.g. $E \subseteq \{(x, y) : (x, y) \in V^2, x \neq y\}$. Network can be directed or undirected, or $E \subseteq \{\{x, y\} : \{x, y\} \in V^2, x \neq y\}$. When the x, y pair is unordered, the network is said to be undirected; otherwise, it is directed. For the rest of the analysis, we focus on simple networks, which disallow multiple edges joining the same pair of vertices.

In many areas that focus on the computation of graphs, adjacency matrices are frequently used alternative representations. An adjacency matrix Y for a simple graph is a square matrix with binary-valued elements y_{ij}. E.g., for an undirected graph, $y_{ij} = y_{ji} = 1$ represents that there is an edge between node i and j and $y_{ij} = y_{ji} = 0$ represents there is not.

1.2.2 Exponential Random Graph Models

The exponential family random graph modeling (ERGM) framework (discussed in detail below) has become a widely used approach for exploring such structure and dynamics of network data [Robins et al., 2005, Lusher et al., 2012]. ERGMs specify the probability distribution for a set of random networks based on exponential family theory. They can also be interpreted as parameterizing a set of local forces that shape the selected micro-structures of the network (e.g., the number of edges, the number of homogeneous connections of an attribute). The advantage of using ERGM is that it follows the basic rules of the exponential family and makes certain numerical calculations simple. ERGMs are widely used in many different research areas, ranging from macro-level (such as country collaboration networks[Nita et al., 2016]), to micro-level (such as Aβ protein networks [Grazioli et al., 2019b]).

Consider a simple undirected network of size N. Its adjacency matrix Y is an $N \times N$ binary-valued random variable describing the state the network. Let y be one realization of Y, the

probability of y can be written as

$$Pr(Y = y|\theta) = \frac{\exp(\theta^\top S(y))}{\sum_{y'} \exp(\theta^\top S(y'))}, \tag{1.1}$$

where θ is the vector of the model parameters associated with the sufficient statistics $S(y)$. The denominator acts as a normalizing factor, making sure this value is a probability:

$$\sum_y Pr(Y = y|\theta) = 1.$$

To use ERGMs to analyze existing network data, a set of sufficient statistics $S(y)$ is defined. Ideally, these statistics are able to fully characterize the focal properties of the said network. Then we can calculate the maximum likelihood estimator of θ given the likelihood function:

$$\mathcal{L}(\theta) = Pr(Y = y|\theta)$$
$$\theta^* = \arg\max_\theta \ \mathcal{L}(\theta) = Pr(Y = y|\theta)$$

1.2.3 Change Scores

The change score describes the changes in network statistics. The traditional description of the change score (e.g. as the *change statistics* defined by Hunter and Handcock [2006], Snijders et al. [2006]), Δ_{ij}, is the difference of network statistics when dyad ij toggles. Let $S(y_{ij}^+)$ be the sufficient statistics of the network when the specific edge y_{ij} is present and $S(y_{ij}^-)$ be that when y_{ij} is not present. The change score Δ_{ij} can be written as

$$\Delta_{ij} = S(y_{ij}^+) - S(y_{ij}^-).$$

A more general definition of change score may also be used. Specifically, $\Delta(y, y')$ is the

difference in the statistics of the two graphs (on the same vertex set, obviously), $\Delta(y, y') = S(y') - S(y)$. The formal definition can be seen as a special case for this, $\Delta_{ij} = \delta(y, y'), y' = y_{ij}^+, y = y_{ij}^+$. This general definition is used in chapter 3 while in chapter 4 we return to the traditional definition.

1.3 Contributions and Outline

In this thesis, we explore three different facets of complex time-varying networks. Chapter 2 examines how commonly used data collection designs affect the dynamics of observed network data, and further impact the imputation performed on such data. Chapter 3 suggests a novel way to approach network dynamics, by investigating the opposite of it, *stability*, to understand the conditions for a network to evolve (or not to evolve). In chapter 4 we propose a novel computational method to the analysis of large scale dynamic networks, which tend to be intractable with the traditional ERGM approaches. We believe that these studies answer various questions about time-varying networks, and provide other researchers with insights and tools to future development of this field.

In chapter 2, we focus on a common data collection method for acquiring network data on longitudinal human interactions - the retrospective life history (RLH) design. While it affords the ability to "peer into the past" vis-à-vis the point of data collection, little is known about the impact of the specific design parameters on the time horizon over which such information is useful. In this chapter, we investigate the effect of two different survey designs on retrospective network imputation: (1) *intervalN*, where subjects are asked to provide information on all partners within the last N time units; and (2) *lastK*, where subjects are asked to provide information about their K most recent partners. We simulate a "ground truth" sexual partnership network using a published model of [Krivitsky, 2012], and subsequently sample this data using the two retrospective designs under various choices

of N and K. We examine the accumulation of missingness as a function of time prior to interview, and investigate the impact of this missingness on model-based imputation of the state of the network at prior time points via conditional ERGM prediction. We quantitatively show that - even setting aside problems of alter identification and informant accuracy - choice of survey design and parameters used can drastically change the amount of missingness in the dataset. These differences in missingness are shown to have a significant impact on the quality of retrospective parameter estimation and network imputation, including important effects on properties related to disease transmission.

In chapter 3 we investigate how, using ERGMs, we could mold the local structure and dynamics of the graph, and further shape the global structure and dynamics of networks. ERGMs can be viewed as expressing a probability distribution on graphs arising from the action of competing social forces that make ties more or less likely, depending on the state of the rest of the graph. Such forces often lead to a complex pattern of dependence among edges, with non-trivial large-scale structures emerging from relatively simple local mechanisms. While this provides a powerful tool for probing macro-micro connections, much remains to be understood about how local forces shape global outcomes. One straightforward question of this type is that of the conditions needed for social forces to stabilize a particular structure: that is, given a specific structure and a set of alternatives (e.g., arising from small perturbations), under what conditions will the said structure remains more probable than the alternatives? We refer to this property as local stability and seek a general means of identifying the set of parameters under which a target graph is locally stable with respect to a set of alternatives. Here, we provide a complete characterization of the region of the parameter space inducing local stability, showing it to be the interior of a convex cone whose faces can be derived from the change-scores of the sufficient statistics vis-à-vis the alternative structures. As we show, local stability is a necessary but not sufficient condition for more general notions of stability, the latter of which can be explored more efficiently by using the "stable cone" within the parameter space as a starting point. In addition to facilitating the understanding of model

behavior, we show how local stability can be used to determine whether a fitted model implies that an observed structure would be expected to arise primarily from the action of social forces, versus by merit of the model permitting a large number of high probability structures, of which the observed structure is one (i.e. entropic effects). We also use our approach to identify the dyads within a given structure that are the least stable, and hence predicted to have the highest probability of changing over time.

In chapter 4 we propose a novel computation algorithm to greatly reduce the cost of applying likelihood calculation on dynamic networks. Currently, Temporal ERGM (TERGM, a variation of the ERGM) requires simulating a network of the entire population, which is computationally intractable on a large scale. For instance, a network with population size n that simulated over time t, the complexity using the TERGM method is $O(n^2t)$. Our approach - dynamic network regression (DNR) has a much manageable complexity $O(nmt)$, where m is the sample size (much smaller than n). In this chapter we investigate the system performance in different settings. We use a synthetic dataset that replicates the properties of retrospective life history (RLH) collected sexual contact network (SCN) data. We examine how different types of missingness introduced by differences in RLH designs affect our system performance. Then we investigate different methods to impute the missing data. We show that with proper RLH designs and with the help of machine learning techniques, our system is able to capture the key SCN properties.

Chapter 5 summarizes my work and contributions, and points out potential shortcomings of some of the methods. This also leads to the future directions that others could follow, to bring forward the research of time-varying networks field as a whole.

Chapter 2

Retrospective Network Imputation from Life History Data: The Impact of Designs

2.1 Introduction

Longitudinal research designs have long been regarded as the ideal data collection frameworks for studying social dynamics [Featherman, 1979]. While prospective designs (which follow a designated set of individuals forward through time) have many well-known advantages, their use is often impractical (or even impossible) when studying phenomena that unfold over long periods of time. In such settings, retrospective designs (which sample individuals at a given point in time and obtain information regarding past events) are an important alternative. For example, fields that study social dynamics over long time spans often employ retrospective life history (RLH) designs [e.g., Forest et al., 1996, Jacobs and King, 2002]. This is due to the fact that, when collecting information regarding the events, relationships, and activities

of individuals over the life course (i.e., *life history data* [Elliott, 2005]), querying a sample of respondents regarding their past experiences is far more likely to be feasible than following a young cohort for multiple decades; and even in the latter case, retrospection must be employed at each interview to obtain events occurring since the last wave of data collection [Scott and Alwin, 1998].

Of the two design types, prospective diary designs have been argued to achieve greater data quality; however, studies have shown that this advantage can be marginal and does not always hold [Hauser et al., 1983, Scott and Alwin, 1998]. This conclusion is particularly important in the settings such as sexual contact network (SCN) studies [e.g. Reading, 1983, Leigh et al., 1998, Tran et al., 2013] where accurately recovering partnership dynamics over long periods can be important for subsequent analysis (e.g., modeling the diffusion of sexually transmitted infections (STIs)). More importantly, diary studies require a high level of subject commitment and ongoing researcher involvement, with concomitant investments of time, money, and effort in recruiting, retaining, and following up with respondents [Stone et al., 1991, Scott and Alwin, 1998, Leigh et al., 1998, Weinhardt et al., 1998]. As such, it is difficult for diary studies to be conducted on large probability samples, or on subjects required to be followed over long spans of time. By contrast, RLH surveys often require only a single interview, have relatively low cost, and can be used in studies that involve larger populations or longer observation periods.

While RLH designs have cost advantages over diary methods, they nevertheless pose significant challenges. RLH studies must focus on significant events that are easily remembered and communicated, due to the fact that 1) retrospective designs are dependent upon subjects' ability to accurately recall life events (often at a substantial remove) [Bernard et al., 1984], and 2) the complexity and length of a life history interview grows rapidly with the number and diversity of past events that are collected [Freedman et al., 1988]. As a result, the need for simplicity and clarity forces researchers to strategically down-sample life events

collected by designs. For example, Add Health: Wave I (1994-1995) asked respondents to list their 3 most recent romantic relationships, and then asked about their sexual behaviors within each of these relationships. In contrast, Add Health: Wave III (2001-2002) only attempted to collect relationships since the last survey (all partners since the year 1995) and surveyed detailed sexual behavior only for the past year [Harris et al., 2009]. Focusing on specific information is necessary for practical data collection; however, each choice of what to measure (and what to omit) brings in some degree of by-design missingness to the data. While the nature of this missingness will clearly vary with the design employed, little is known about how RLH designs induce missingness in network data, or what impact such missingness would have on retrospective inference. In this chapter, we provide the first quantitative investigation of this issue.

We note from the outset that missing data in longitudinal studies can accrue through multiple mechanisms. De Leeuw [2001] provides a review of the types of missingness that commonly appear in conventional survey data. For instance, item non-response is a source of missingness that is relatively well studied. Loss of information at the item level may occur due to information not being provided by the subject, provided information being unusable, or usable information being lost, all of which are examples of unintended or *out-of-design* missingness. Little and Rubin [1987] studied many of the mechanisms behind item non-response, and proposed targeted strategies based on whether items are missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR). Morris [1993a] studied this phenomenon in the SCN context by examining the discrepancy between male and female reported sexual partners; more broadly, Smith and Moody [2013] and Smith et al. [2017] use simulation studies to examine various sources of missingness that affect cross-sectional network studies, with an emphasis on identifying the robustness of descriptives to data loss. Unlike these other sources of missingness, the *by-design* missingness arising in RLH studies is not well characterized. In particular, little work has been done on missing spell data caused by design in retrospective longitudinal studies, particularly when such spell

data encodes relational structure.

In this chapter we focus exclusively on by-design missingness (i.e. the data that is *necessarily* lost because of the questions that were or were not asked, as opposed to data that *could be* lost, depending on subject behavior). Likewise, in some studies the population of potential actors is known (or can be closely approximated), while in others this too must be inferred from the data (leading to the possibility of missing individuals, or *vertices*, as opposed to only missing relationships, or *edges*). Here, we focus on the problem of missing edges, which arises even when the population is fully characterized. As we show below, common study designs introduce complex and consequential patterns of missing data even under these optimistic conditions.

In the remainder of this chapter, we formally characterize and analyze two common types of designs that are widely used in RLH studies, which we dub *intervalN* and *lastK* designs. We will focus on how information decays in these two designs as one moves from time of interview into subjects' relative past, and the impact of this information loss on inference. We first examine general properties of information loss through these two design families, including asymmetries in loss of information for existing relationships (edges) versus non-relationships (nulls), which are broadly applicable to any RLH network study. We also show that *lastK* designs based on ordering of the end times of spells (terminal selection) have generic advantages over designs based on initiation of spells (onset selection) no matter what the underlying data, leading to a general recommendation over the former. To provide a more quantitative sense of how by design information loss might play out in practice, we further perform a simulation study using synthetic SCNs based on a large data set (the NHSLS [Laumann et al., 1996]). We start by examining the simple quantity of missingness introduced by each family of designs as a function of time before interview, and the extent to which this missingness is focused on existing relationships (edges) versus non-relationships (nulls). To get a better sense of how this missingness translates into loss of inferential power,

we examine the impact of accumulated missing data on a "Drosophila task," namely inference for cross-sectional exponential random graph model (ERGM) [Robins et al., 2007] parameters fit to pre-interview periods (comparing results from sampled data with corresponding complete-data results). Lastly, we translate these inferential effects into pragmatic terms by examining the effect of missingness on our ability to impute network properties at previous time points. As elaborated below, SCNs are a natural choice for such a case study, being substantively important, frequently measured using RLH designs, and relatively well-modeled using existing techniques. However, the insights from these analyses generalize to many other types of spell data collected with retrospective designs, and we discuss some of these implications throughout the chapter. Finally, we close with some observations and recommendations vis-à-vis retrospective life history designs for retrospective inference on sexual contact (and other) networks.

2.1.1 Sexual Context Networks as an Important Case

While, as noted, we use sexual contact networks as a case with which to consider broader questions regarding retrospective life history designs, it should also be noted that this case is sociologically important in and of itself. SCNs lie at a complex intersection of social, cultural, economic, and biological phenomena: sexual behavior is a vital dimension of human activity that is interwoven with other social institutions and economic arrangements (marriage, cohabitation, sex work, etc.), that is linked with life course transitions such as the transition to adulthood [Sassler et al., 2018], and that has obvious and significant implications for socially relevant biological phenomena such as fertility [Guzzo, 2014, Everett et al., 2017] and disease transmission [Robinson et al., 2013, Doherty et al., 2005]. The structure of SCNs has been implicated in systematic health disparities affecting vulnerable groups [see e.g. Adimora and Schoenbach, 2005, Hamilton and Morris, 2015, Mustanski et al., 2016], most notably substantial differences in HIV prevalence, with implications that carry over

into economic disadvantage, early mortality, and social stigma [Pellowski et al., 2013]. SCN structure has even informed health policy debates (e.g., about the effectiveness of targeted versus broad-based interventions for combating sexually transmitted infections [Wilson et al., 2008, Morris et al., 2009], the potential value versus intrusiveness of attempted cultural interventions (e.g., Uganda's well-known "zero-grazing" campaign [Green et al., 2006]), and the regulation of sex work [Hsieh et al., 2014]). Given the wide range of social phenomena to which they are mechanistically related, SCNs are an important target for sociological investigation.

Other reasons for using SCNs as our working case in the context of this chapter are more pragmatic. Most obviously, retrospective life history designs are an important means of collecting SCN data, so there is a natural fit between the case and the broader issues of RLH designs. SCNs have also played an important role in the development of statistical network modeling [see e.g. Morris, 1991, Morris and Kretzschmar, 1997, Hamilton et al., 2008, Krivitsky et al., 2011, Krivitsky, 2012, Carnegie et al., 2015], and they are hence a natural case to consider from that point of view; relatedly, they are currently unique in having been used to create data-calibrated population-level network models that can be used for realistic synthetic data studies. Finally, SCNs are relatively simple in the sense of being highly sparse with relatively weak dependence and low clustering compared e.g. to friendship networks, which makes them a good "base case" for an initial study of a potentially complex problem (compare with e.g. their use by Butts [2011] to study network asymptotics). In this spirit, we emphasize that the general approach to codifying RLH designs introduced here, and many of the additional results, apply to contexts far beyond the SCN case, though that case is in and of itself an important one.

2.2 Background

In this work, we employ the following terminology in describing the measurement process, data, and analysis. A participant or subject of a survey study is referred to as an *ego*, and each of his/her connected individuals (e.g. friends in friendship network studies, or sexual partners in a sexual contact network) is called an *alter*; their attributes are called *ego* and *alter covariates*; a relationship is referred to as a *tie* or a *spell*; the start and end times of a spell are called respectively the *onset* and *terminus* (plural: *termini*). The collection of data from the respondent is generically referred to here as *measurement* (whether conducted via an interview, self-administered survey, or other approach) and the *measurement time* is the time at which the data is collected. A *measurement interval* refers to the interval of interest a design covers. In what follows, we will limit ourselves to the case in which all respondents are measured at the same time. An assessment of the (cross-sectional) state of the evolving network at a given point of time is referred to as a *query*, with the *query time* being the time point in question. Obviously, we are interested here in *retrospective queries*, for which the query time precedes the measurement time; this time difference is referred to as the *look-back time*, and the data associated with such a query (i.e., the known and missing edge states) is referred to as *retrospective data*. Unless noted otherwise, we describe the measurement time as the "present," with queries at higher look-back times delving further into the (relative) "past." Intuitively, our objective is then to understand how missingness and information loss accrue with look-back time, as a function of the retrospective life history design being employed.

2.2.1 Survey Designs in a Retrospective Life-history Context

An RLH measurement can be viewed as a two-stage sampling process. First, survey respondents are sampled from a broader population. Second, their life events (whether unique oc-

currences or partnership onset/termination events) are then sampled from the pre-interview history of each subject. Schematically, we may view this process in terms of a *life event space* as shown in fig. 2.1; the x-axis represents time relative to interview, and the y-axis indexes members of the population. Each horizontal line represents one individual's (partial) life span and the points are his or her life events. This figure illustrates two-stage missingness of retrospectively collected longitudinal data: Viewing from the vertical axis, we only observe the events lying on the solid lines (sampled individuals); while viewing from the horizontal axis, our choice of RLH design determines which events (of the sampled individuals) we are able to observe and which are missing. We note that the spell data can also be illustrated on this life event space, where the onset and terminus of a spell are plotted as two events, and the segment in between denotes the spell.

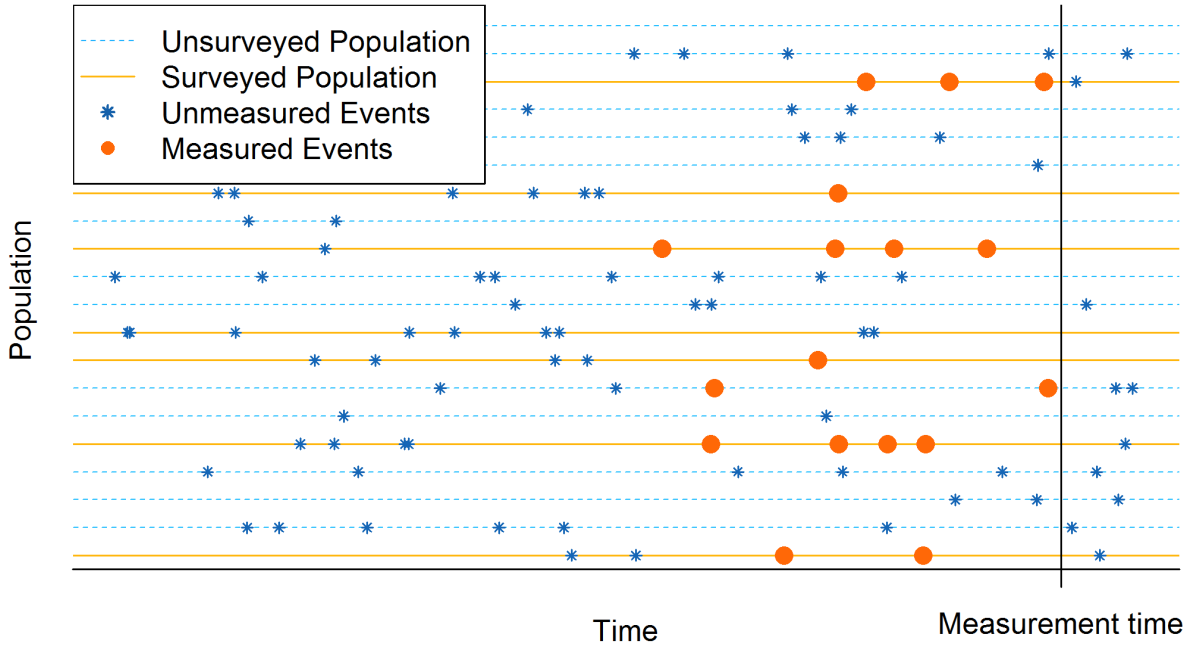

Figure 2.1: A schematic illustration of an RLH survey in an event space. The x-axis represents time relative to interview, and the y-axis indexes individuals in the population of interest; horizontal lines reflect individual histories, with dots and stars representing life events. Differences in RLH designs affect not only the individuals whose events are observed, but also the events that are seen within any given window of time.

To study longitudinal spell data, researchers rely on RLH designs to collect a range of

information regarding both individual behaviors and the evolving network structure. Some of the most commonly collected types of information include the ego and alter covariates and the timing of their spells. In the specific case of SCN studies, researchers are in some cases focused on *egocentric* questions involving the behavior of individuals and their relationship to the local social environment, while in other cases researchers studying SCNs are interested in the global structure of the evolving contact network. While studying the latter from egocentric data was long viewed as infeasible, recent advances in statistical network modeling [Krivitsky, 2012] have brought this goal within reach. Although the impact of cross-sectional sampling (i.e., of egos) on network inference has been studied [Handcock and Gile, 2010, Krivitsky et al., 2011], the implications of RLH designs on missingness *through time* has not. For instance, imagine our task is to make a query to the network formed by the respondents shown in fig. 2.1 at a particular time, to extract an ego's life events and network structure as a whole. The events we are able to observe depend on which events are collected by the design, and the design will hence influence the conclusions we draw regarding the state of the network at that moment.

In this chapter we study two commonly used design types in RLH surveys. In the first design, egos are asked to provide information on all alters over a specified time window of length N; in the second design, egos are asked to provide information on their last K alters. We refer to these two design classes respectively as *intervalN* and *lastK* designs. For the purpose of our analysis, we assume that the measurement interval in the intervalN design ends with the interview time (making it comparable to lastK). The parameters of each design are defined as follows: The intervalN design can be (1) *uncensored*, in which we record the start or end time of a relationship, even if it starts before or ends after the measurement interval, as long as a part of the relationship overlaps with the interval; or (2) *left/right censored*, in which the start time (left-censored) or the end time (right-censored) of a relationship is not recorded if it is outside of the collection interval. It is worth noting that many RLH datasets are right censored because the termini of active ties are unknown at the time of

survey and additional follow-ups are not performed to determine when the ties (eventually) end. Thus, RLH data collected with a single interview is almost always right-censored, unless all on-going ties are discarded from the dataset. There are two basic types of lastK designs, differentiated by how "last" relationships are defined: (1) *onset selection*, in which subjects are asked to provide information on the K relationships having the most recent start times; or (2) *terminal selection*, in which subjects are asked to provide information on the K relationships with the most recent end times (with ongoing relationships being counted first). It should be noted that, while our focus is on the use of these designs in network studies, these designs can be employed when eliciting spells of any sort; thus, many of our observations regarding their basic properties are broadly applicable to more general RLH studies.

These two types of designs are common in retrospective life history studies, particularly for SCNs, with many data collection efforts adopting designs that are either lastK, intervalN, or some combination of the two. For instance, the Add Health dataset [Harris et al., 2009] used multiple designs in different waves. Add Health: Wave I (1994-1995) asked respondents to list their 3 most recent romantic relationships, while Add Health: Wave III (2001-2002) asked about detailed sexual behavior only for the past year. The National Health and Social Life Survey [Laumann et al., 1996] collected subjects' detailed sexual behavior within one year from the interview time, and asked the subjects to list all sexual partners within that time period (up to a maximum of 28); since complete information was obtained on all spells collected, this design is of a left-uncensored intervalN type. An interesting instance of a "firstK" design (asking for the first K) partners can be found in the first British National Survey of Sexual Attitudes and Behaviors [NATSAL-I; Johnson et al., 1994]; NATSAL-II [Erens et al., 2001] and NATSAL-III [Erens et al., 2014] employed lastK designs, in addition to broader questions on numbers of sex partners and related information. Implicit designs that are equivalent to lastK or intervalN can also arise in e.g. online social network (OSN) studies, due to query restrictions imposed by site operators. For instance, Mayer and Puller

[2008] studied Facebook friendship ties using data from 2005, while Facebook was only launched in 2004 (IntervalN, left-censored).

Supplementary fig. C.1 provides a schematic illustration of how the two designs sample events differently within the life event space in a hypothetical case, highlighting the contrast between lastK and intervalN designs. While many relationships would be caught by both designs, there are systematic differences in which relationships are captured. As we shall show, these differences have substantial implications for the researcher's ability to infer the state of a network over time.

2.2.2 Relational Missingness in lastK and intervalN Designs

Even in the special case in which all members of a population are interviewed (a network census), RLH designs impose constraints on the relationships that can be observed. For instance, Figures 2.2 and 2.3 illustrate the patterns of spell-wise missingness that result from intervalN and lastK designs, respectively, in a hypothetical case involving a respondent with five relationships prior to time of interview. Intuitively, most designs prioritize spells near the measurement (interview or survey time); nevertheless, the exact spell selections vary markedly.

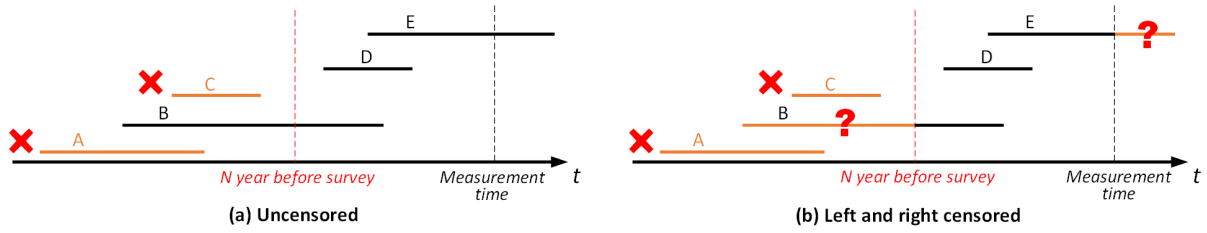

Figure 2.2: Missingness in intervalN designs. A hypothetical life history is shown with spells (A-E) prior to interview time. In both censored and uncensored cases, spells that lie entirely outside of the measurement interval are not observed. Ties B and E lie partially inside of the measurement interval, their onsets and termini are observed on uncensored case (a) but are not observed in the left and right censored case (b). In (b), only portions of spells overlapping with the measurement interval are observed.

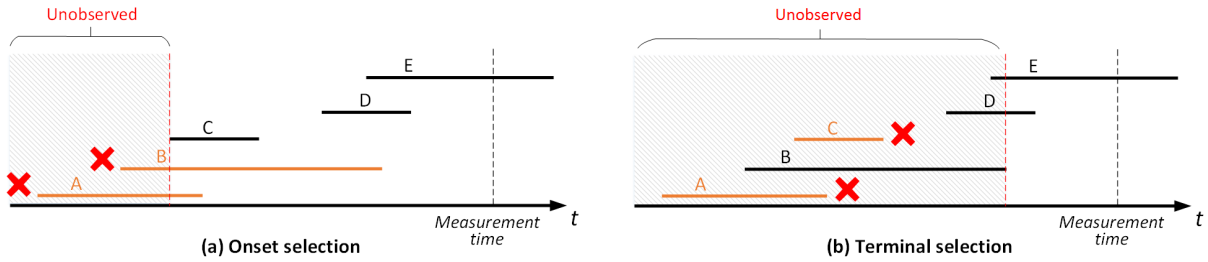

Figure 2.3: Missingness in lastK designs. A hypothetical life history is shown with five spells (A-E) prior to the time of interview. (a) Observed and missing spells for a lastK design with $K = 3$ in onset selection: any ties that start before the onset of the 3rd tie (tie C) will not be recorded, and thus ties A and B are unobserved. (b) terminal selection: ties ending before the terminus of 3rd tie (tie B) are not recorded, and thus ties A and C are unobserved.

These patterns of missingness have important implications for retrospective network inference - that is, our ability to recover information on the state of the network prior to the measurement time. Taking the above scenario as an example, Figures 2.4 and 2.5 respectively illustrate the implications of intervalN and lastK designs for inferring network structure. We make retrospective queries to obtain each ego's ties at particular times prior to measurement (red and blue vertical lines) to collectively reconstruct the network state. We express his or her relationships in the then-current network adjacency matrix. Edges and *nulls* (i.e., non-ties) are respectively represented by 1's and 0's; the unobserved edges and nulls are called *missing edges* and *missing nulls*, and are represented by question marks. (Note that a missing edge is not equivalent to a null, as the former refers to an unobserved edge while the latter refers to a potential edge observed to be absent from the network.)

Arguably, the simplest pattern of retrospective cross-sectional missingness arises in the intervalN case. Within the measurement interval, we have full information on ego's personal network, and are able to recover both ties and nulls. Outside of this interval, the pattern of missingness depends on the design censoring scheme. As illustrated in fig. 2.4, when censoring is present, onsets or termini will be missing if a tie crosses an interval boundary; we do know, however, whether the tie was established at or before the beginning of the interval or has not yet finished at the end of the interval. Ties not extending into the sampling inter-

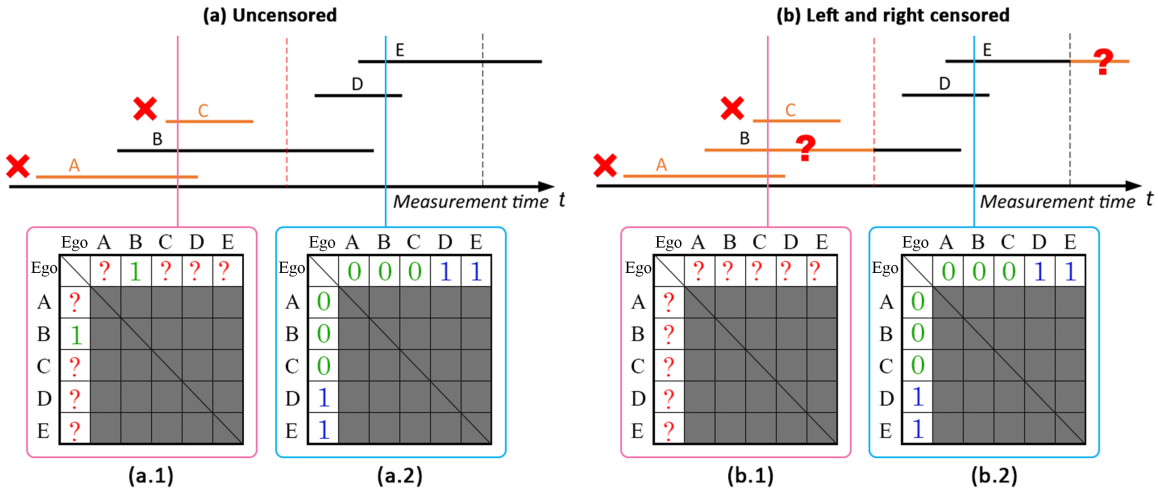

Figure 2.4: Cross-sectional missingness of the intervalN design illustrated with adjacency matrix (first row/column indicates ego). (a.1) and (b.1) are the adjacency matrices of the network cross-sections taken from a time slice before the measurement interval. Only (a.1) was able to capture ego's tie to B (a.2) and (b.2) are taken from within the measurement interval and all ties are observed.

val and nulls outside the interval are both unobserved for time points outside the interval, yielding little additional information.

By contrast, the pattern of missingness induced by lastK designs is more complicated. Fig. 2.5 shows a few common scenarios. In onset selection and terminal selection, unreported ties can have different implications. In onset selection, if a tie has been observed in earlier time slices, we know that no additional spell involving a tie to the same actor can be observed a later period. In terminal selection, if the query time is after the terminus of the Kth tie, then all unobserved ties are nulls.

Figure 2.5 panel (a) shows a scenario involving a last 3 partner design using onset selection. C is the third most recent tie based on the start time and we know there is no other tie that starts after the onset of C. However, there could exist ties that starts before the onset of C but and stayed active longer, e.g., B. At the first query point (a.1) C is the only tie that is observed, and any other edges or nulls are unobserved. At the second sampling point (a.2), because we have previously observed C at (a.1), we know there is no other tie

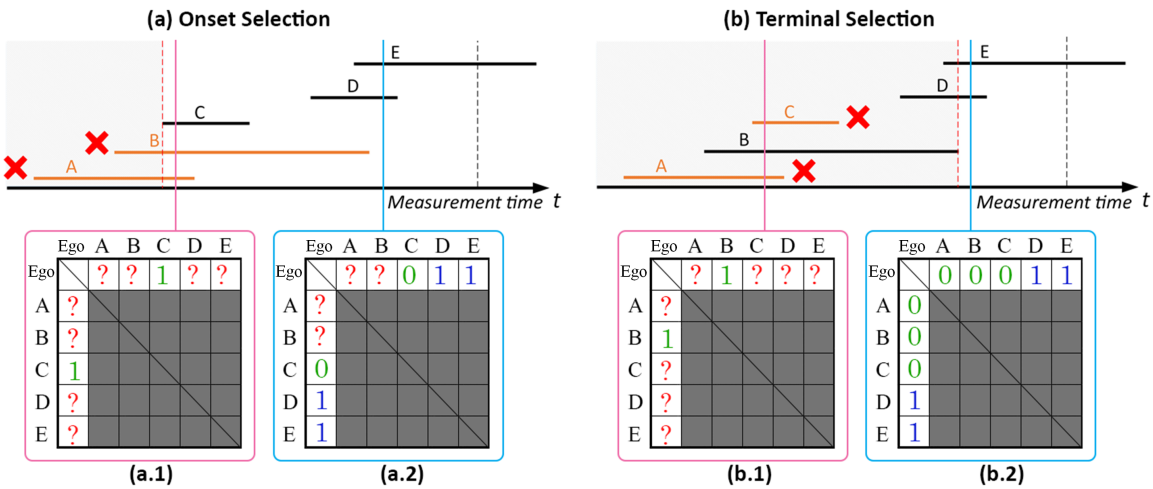

Figure 2.5: Cross-sectional missingness in the lastK design illustrated by the adjacency matrix (only the row and column that are associated with ego are shown). (a.1) and (b.1) are the network cross-sections taken from a time slice relatively far prior to the measurement time; while (a.2) and (b.2) are taken from time slice that is near the measurement time. In both (a.1) and (b.1), we are uncertain about whether there are previous ties to D and E occurring prior to the observation (e.g., episodic or reoccurring relationships). In (a.2), we are certain that a tie to C is not present because if there had been a tie to C after the previously observed C tie, onset selection would have sampled it instead. In (b.2), we are certain that ties to $A - C$ are not present, because otherwise terminal selection would have sampled them instead of the tie to B.

C' with the same partner that starts after C has terminated (because otherwise C' would have been picked up instead of C). This pattern is a marked contrast from the pattern associated with an equivalent lastK design using terminal selection (panel (b)). Here, we have full information on all edges and nulls from the end of the Kth edge forward in time (since any edge extending past this point would have itself become the kth edge). Prior to this, our information is more limited; for instance, at the red query point, we can observe only the presence of the edge to B, with all other edges and nulls being unknown.

2.3 A Simulation Study of RLH Design Missingness

While all of the above designs tend to preserve more information for query times closer to the measurement time, missingness accumulates in distinct ways. This raises the question of how information is lost, and what are the impacts of different design choices as we look back in time. Given the need to query the network at a particular time point, we would ideally like to be able to anticipate how much data will be observable (and the nature of what is lost). Likewise, we would like to be able to assess the consequences of such missingness for retrospective inference, and to employ this information to inform our choice of study design. We approach these problems via a simulation study in which we begin with a known *ground truth* network and then apply either intervalN or lastK designs to probe the incidence and consequences of missing data. For reasons noted in Section 2.1.1, we focus on the case of sexual contact networks, using simulated networks based on the NHSLS as our inferential targets.

We employ the following study procedures (illustrated in fig. 2.6) to quantitatively assess missingness and the impact of the RLH designs. First, we generate a longitudinal ground truth network (as described in section 2.3.1), expressing it as a time series of cross-sectional networks. We call these cross-sectional networks the *true networks*, as they contain all partnership information at each time point. Then we simulate an idealized RLH interview process on each of the true cross-sectional networks to get a series of the *observed networks*, with NAs representing unobserved ties. We then apply three evaluation metrics (described in section 2.3.2) to the true and observed networks in order to assess the impact of the RLH design on retrospective inference. While, as in any simulation study, the networks we employ do not reflect the full complexity of networks observed in real-world data, they do provide a starting point for understanding how missingness would be expected to accrue in RLH designs. As we show, many aspects of information loss from intervalN and lastK designs follow basic patterns that are likely to generalize not only to real-world SCNs, but to other

types of networks as well.

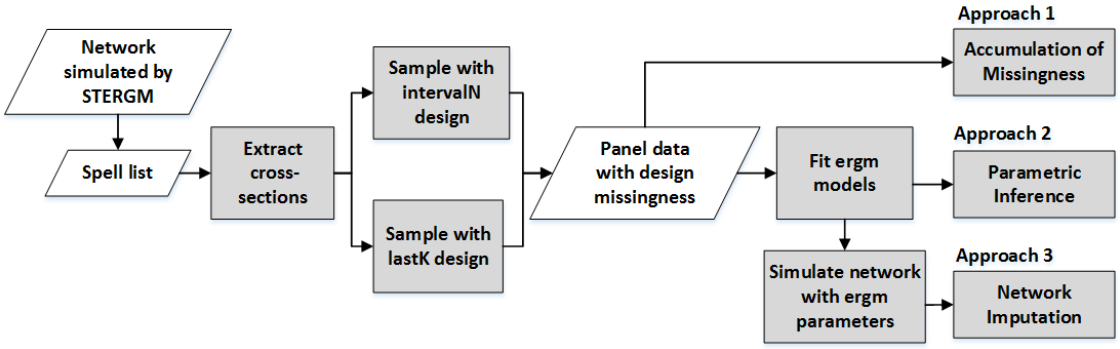

Figure 2.6: Schematic workflow for the simulation study. This process was replicated 600 times to generate the study data set.

2.3.1 Simulation Model

In order to assess how missingness accrues under different designs, "ground truth" networks without design missingness are required. Ideally, we would employ a real world sexual contact event dataset that contains complete partnership information with no missingness over a long period of time, so that we could perform intervalN and lastK sampling with realistic values of K and N without encountering data limitations. Unfortunately, all existing SCN datasets lack the quality and coverage to be used in this capacity: many studies only contain the detailed spell information for a short period of time, and almost all studies suffer from subject/item non-response missingness, which makes it difficult for our study to isolate the by-design missingness. Given these limitations, we instead employ synthetic data generated from a previously published dynamic network model of Krivitsky [2012] (based on data from the NHSLS [Laumann et al., 1996]) which has been shown to reproduce many properties of real SCNs. This model belongs to the Separable Temporal ERGM (STERGM) family [Krivitsky and Handcock, 2014], which models tie evolution of social relationships via the interaction of a formation process (in which new ties are generated) and a dissolution process (in which old ties are broken). The model (whose parameters were inferred from

egocentrally sampled SCN data from the NHSLS) includes effects for gender, race, and age mixing, relationship duration, concurrency, as well as other factors found to influence SCN dynamics, and was estimated at monthly time resolution [Krivitsky, 2012]. Given this fitted model, we may simulate complete synthetic SCN data for a hypothetical population with realistic dynamic and cross-sectional properties; sampling from this data (and performing retrospective inference) thus allows us to examine the impacts of RLH designs *per se*, above and beyond other factors that can affect data quality.

For our simulation study, we draw 600 synthetic populations of size 500 (with replacement) with individual demographics chosen to match randomly selected respondents from the NHSLS sample using their sampling weights. Given each synthetic population, we simulate a SCN for 500 monthly time steps (following a 200 month burn-in period); for simplicity (and because the STERGM model does not support demographic change), we employ a "static population" approximation, with individual characteristics fixed over the simulation period. Although nodes entering and leaving the network is an important factor in many study settings, this phenomenon does little to impact the outcomes of our particular study.[1] Descriptive analyses confirm that the simulated SCNs reproduce key features of the NHSLS networks. In particular, the average tie duration is around 9 years (106 months), the average degree of the network is 0.74, and the concurrency rate is 0.17. We compare these properties with measures from NHSLS: the average tie duration is 108 months; the average degree is 0.71; and the concurrency rate is 0.06. The close match between our synthetic networks and the NHSLS suggest that the former are a reasonable proxy for experimental purposes.

In order to focus on the impact of RLH design choices *per se*, we here assume that survey design is the only source of missingness in the datasets, eliminating out-of-design missingness mechanisms such as interviewer error or informant non-response. For the same reason, we

[1]Nodes that have departed would not be subject to measurement under either design, and vertex missingness is not a target of interest here would not be available to be queried); nor does the entry of new nodes pose special challenges. Demographic change could alter the detailed structure of the network somewhat, but our simulation nevertheless reproduces the key features of the NHSLS that are important for our application.

treat all members of the synthetic population as respondents (i.e., we simulate a network census study). Our scenario thus intendedly represents a "best case" for information that can be obtained from a given design, allowing us to identify hard limits on what can be inferred under more realistic conditions. With these ground truth networks, we then decide which ties/nulls are fully or partially observed based on intervalN or lastK designs with choices of N and K.

2.3.2 Evaluating the Impact of Designs

One challenge in evaluating complex designs is determining the effective extent and nature of information loss associated with in-design missingness; this is particularly true in the context of relational data, where dependence among observations and the nonlinear nature of most estimands of interest make intuition an unreliable guide. Here, we address this issue through three linked approaches that allow us to both describe the extent of missingness associated with particular designs and to trace out the implications of such missingness for subsequent analyses:

- *Accumulation of Missingness:* This is the most fundamental assessment of network missingness. We calculate the accumulation of missing edge and null variables at query times in the increasingly distant past (relative to the time of interview). This provides us with a simple description of how the fraction of known states within the network adjacency matrix declines with retrospection, an important factor in subsequent analyses.

- *Parametric Inference:* The extent of information loss due to missing data may depend on *which* edge variables are missing, as well as the amount of missingness. To assess the loss of information as we recede into the past, we examine the effect of missing data on our ability to correctly infer the parameters of a cross-sectional ERGM fit to

observations at various time points prior to the interview time.

- *Network Imputation:* While declining accuracy in parameter estimates provides one indicator of information loss, this does not immediately translate into the past states of the network. To capture this, we also perform model-based imputation of the state of the network at each query point, allowing us to assess how knowledge of network structure *per se* declines as we move into the past.

Implementation of the first approach is straightforward: at each query time, we calculate the number of edges and nulls that are missing due to the design. We then examine the trend of the missingness in the networks as a function of time prior to the measurement time (i.e., look-back time). The second approach uses an ERGM fit as a standardized tool to assess information loss. At each query time, we fit the retrospective data with a basic ERGM model with effects that are related to the dynamic model and are common in SCN studies (shown in 2.1), thus emulating a typical analysis that might be conducted in the absence of dynamic data. We also fit the identical model to the ground truth data, providing a set of "true" parameters (i.e., the parameters that would be obtained from such an analysis in the absence of missingness). We then examine the absolute differences of the ERGM coefficients drawn from the true and retrospectively observed cross-sectional networks at each query point. Finally, our last approach focuses on the impact of missingness on retrospective network imputation. For each observed cross-sectional network, we use the fitted ERGM parameters from the previous approach to impute the network state via conditional simulation where the observed portion of the network is held unchanged and the missing dyads are imputed [Wang et al., 2016]. From this we calculate a number of basic structural properties that are common in SCN analysis, such as degree distributions and diffusion potential. We then compare these estimated properties with those of the ground truth network. Per Fig. 2.6, each of the above analyses is conducted on each replicate simulation in each condition, yielding a distribution of outcomes for each query time and design. (Model estimation and simulation were both

performed using the `ergm` package for R [Hunter et al., 2008] with additional calculation of network properties performed using `network` [Butts, 2008b] and `sna` [Butts, 2008c].)

Table 2.1: List of terms used in the cross-sectional ERGM.

ERGM terms	Descrption
`edges`	Intercept
`nodematch('sex')`	Gender homophily
`absdiff('age')`	Age homophily
`degree(1)`	Tendency towards monogamy
`nodematch('ethcat')`	Race/ethnicity homophily

2.4 Results

2.4.1 Accumulation of Missingness

IntervalN Design

We start by showing how missing edges and nulls accrue under an intervalN design. In the censored case, where both the onset and terminus of a tie are not recorded if it lies outside of the length-N interval, the pattern of missingness is very straightforward: within the measurement interval, all edges and nulls are recorded; otherwise, none of them are recorded. This gives missingness a binary behavior: we either observe everything or we observe nothing, depending on whether the network cross-sections lie within the measurement interval.

As mentioned in section 2.2.1, most RLH designs are right censored (and left uncensored), and the missingness of this case is illustrated in fig. 2.7. In this plot we show four different N values, $N = (1, 12, 120, 240)$ months. Similar patterns are displayed for all N's: edges and nulls are fully observed within the length-N measurement intervals. As we look back in time, once we pass the beginning of the measurement interval, we immediately miss *all* nulls

and most of the edges, except for those edges that overlap with the measurement interval, hence causing the sharp rise in missingness. As we move further into the past passing onsets of all observed ties, the proportion of missing data reaches 1.

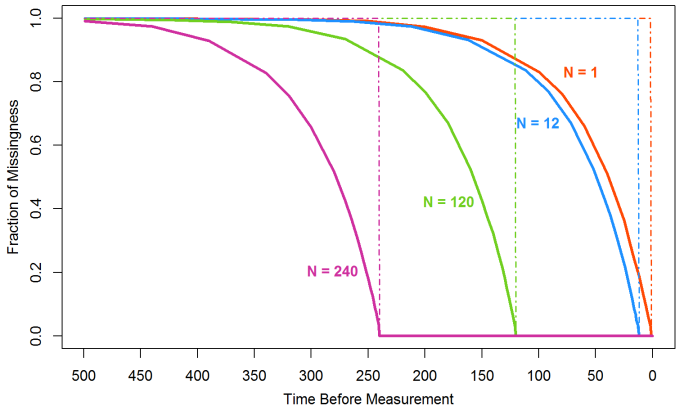

Figure 2.7: Accumulation of missingness under intervalN designs. Horizontal axis indicates look-back time, while the vertical axis indicates fraction of data missing (by type). Results are shown for designs with Ns of 1, 12, 120, and 240 months.

	Fraction of missing ties					
	Before measurement			Before beginning of MI		
N (yr.)	0.25	0.5	0.75	0.25	0.5	0.75
$^1/_{12}$	1.4	3.3	6.5	1.3	3.2	6.4
1	2.3	4.2	7.4	1.3	3.2	6.4
10	11.2	13.2	16.4	1.2	3.2	6.4
20	21.3	23.2	26.5	1.3	3.2	6.5

Table 2.2: Number of years it takes to miss X fraction of the total ties for the intervalN design. The years are calculated from both the interview time and the beginning of the measurement interval (MI). The 4 N values are taken from the figure above (fig. 2.7).

LastK Design

Fig. 2.8 shows the missingness of two lastK variants: (a) terminal selection and (b) onset selection. We experiment with four different K values, $K = (3, 4, 5, 6)$, which are commonly used in sexual partnership surveys (e.g., [Harris et al., 2009]).

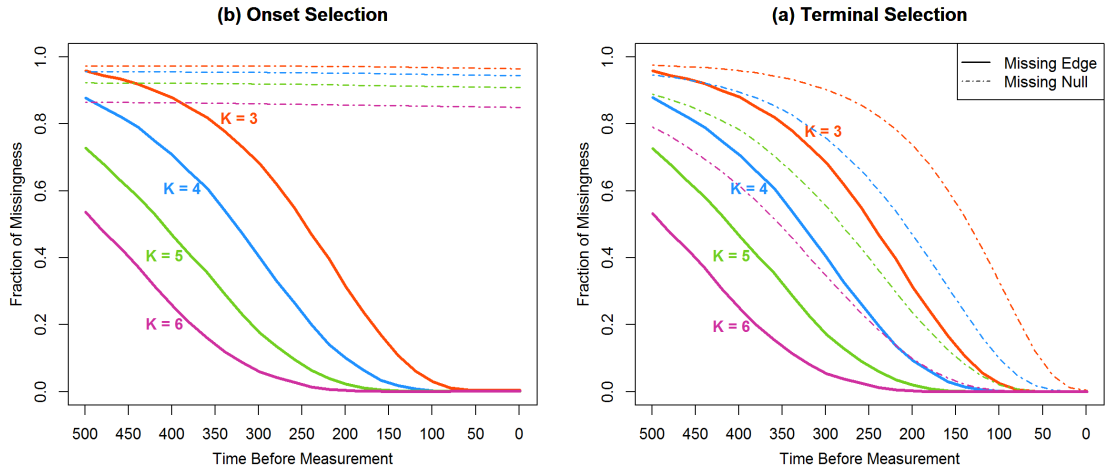

Figure 2.8: Accumulation of missingness under lastK designs. Horizontal axis indicates look-back time, while the vertical axis indicates fraction of data missing (by type). Results are shown for onset selected (a) and terminal-selected (b) spells with K values of 3–6.

	Fraction of missing ties					
	Terminal selection			Onset selection		
K	0.25	0.5	0.75	0.25	0.5	0.75
3	11.9	18.2	25.4	15.3	20.6	27.3
4	15.9	23.4	31.7	21.3	27.3	35.1
5	19.8	28.1	38.4	27.2	34.3	-
6	24.1	33.1	-	33.1	40.6	-

Table 2.3: Number of years it takes to miss X fraction of the total ties for the lastK design family. The years are calculated from both the interview time and the beginning of the measurement interval. The K values correspond to fig. 2.7.

1. Onset Selection: i) As we look back in time prior to the survey, the missing fraction of both ties and nulls gradually increases from 0 to 1, following a sigmoidal function in look-back time. ii) By contrast, the fraction of missing nulls is stable and close to 1 at all times for all K values. This is because selecting the most recent K ties based on onset does eliminate the possibility that there could be a long-lasting tie that started before the onset of all K ties and is still active at the time of measurement. The missing fraction is not exactly 1 because nulls associated with persons having fewer than K partners can be observed (see Fig 2.5.a). iii) A larger K value corresponds to a slower increase of missing tie fraction as a function of look-back time. It also corresponds to a lower missing null fraction.

2. Terminal Selection: i) Compared with onset selection, a similar sigmoidal function of look-back time is displayed with missing ties. ii) Missing nulls accumulate much faster than the missing edges. To explain this phenomenon we make reference to fig. 2.3.b: before the end of the Kth tie (tie B in this case), we do not observe any nulls (A and C) but we could observe multiple edges (B, D and E) that start before and end after the end of the Kth tie. iii) A larger K value corresponds to a slower increase of both missing ties and missing null fraction as a function of look-back time.

Advantages of Terminal vs. Onset Selection

There is little extant literature differentiating onset selection and terminal selection in lastK designs. Our simulation results show that the terminal selection design holds significant advantages in capturing missing nulls with a similar performance on missing ties. To further characterize this phenomenon, we prove in the appendix (section A) that if two subdesigns pick different sets of ties, the ones picked by terminal selection always have longer or equal durations (i.e. ties cover equal or longer windows of time). Furthermore, we are able to show that the ties picked by terminal selection (if different from onset selection), always start no later than, and end no earlier than the ties picked from onset selection.

The above observation suggests a strong advantage to using terminal selection. Measuring longer ties generally reduces the number of ties missing in cross-section. Observing ties that end more recently also suggests better data quality if the researcher is more concerned with the subjects' current status. Given that the terminal selection approach has clear advantages with few obvious weaknesses, we recommend its use as a default lastK design unless circumstances dictate otherwise. For this reason, our remaining analyses on lastK employ the terminal selection variant.

2.4.2 Impact of Designs on Parametric Inference

IntervalN Design

The parametric inference results for the intervalN design can be immediately deduced from the pattern of missingness. Unlike the lastK design, where the missingness accrues gradually from the time of measurement, intervalN displays a sharp change in available information at the boundary of the measurement interval. When inside the measurement interval, ties and nulls are completely observed (implying no impact on inference). However, once we step outside of the measurement interval, the missingness rate for edges grows rapidly and the nulls are completely unknown (see fig. 2.7). Thus, the intervalN design imposes no degradation of inferential quality vs. complete data for look-back times within the measurement interval, while inference beyond this interval is essentially impossible (since no nulls can be observed, the likelihood provides almost no information on density).

LastK Design

Fig. 2.9 illustrates the absolute difference in ERGM parameter estimates for complete data versus sampled networks as a function of look-back time. Consistent with our findings in overall levels of missingness, we see that inference degrades with look-back time and improves with increasing K. Interestingly, inferential performance does not always degrade smoothly, with quality tending to be very high over short to moderate look-back times before undergoing a sudden transition to instability. Performance outside this "safe" interval can be erratic, with some samples yielding high-quality estimates and others being quite poor. This is especially true for parameters associated with relatively rare events. For instance, the parameter for same sex ties often diverges at longer look-back times (e.g. produces more infinite estimates), a consequence of no same sex ties being captured within some samples.

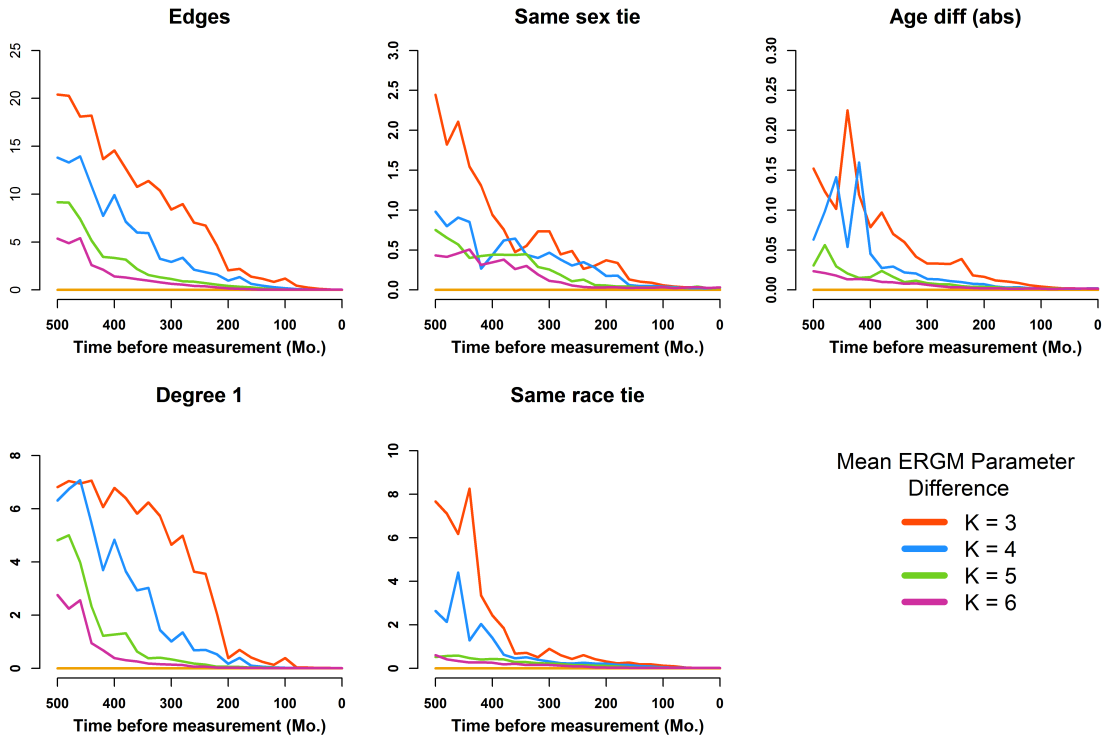

Figure 2.9: Effect of lastK design on parametric inference. Solid lines show the mean absolute difference in estimated ERGM parameters for true versus observed networks over 50 simulations. Each subplot shows errors as a function of look-back time for a given model parameter, for each simulated K value.

This, in turn, contributes to instability in other parameter estimates. Fig. C.2 (in Appendix) provides a more detailed depiction of the distribution of outcomes in each case.

2.4.3 Impact of Designs on Retrospective Network Imputation

We use two different measures to examine the effect of design choice on network imputation. First, we compare the degree distributions of the true network and the conditionally imputed network. We select (simultaneous) degree as a key measure because many epidemiological studies argue that concurrency (characterized by having ties with 2 or more partners simultaneously) is one of the most important factors driving STD prevalence in SCNs [Morris and Kretzschmar, 1997, Rosenberg et al., 1999]. As a second measure, we also examine the av-

erage number of reachable vertices at a given time point. This is also an important measure for assessing the network robustness against STI. Assume we randomly pick one individual from the cross-sectional network and mark him or her as "infected" with a hypothetical rapidly transmitted STI. We are then interested in knowing the number of people this disease can potentially reach in a short time period (i.e., before the network evolves further). This corresponds to a weighted average of the instantaneous component size distribution, with weights being proportional to the number of people in each component. Formally, let the graph at time t be composed of components $C_1^t, \ldots, C_m^t$ with respective sizes $|C_1^t|, \ldots, |C_m^t|$; the expected number of instantaneously reachable vertices from a random seed node is then

$$\mathbf{E}R_t = \frac{\sum_i |C_i^t|^2}{\sum_i |C_i^t|}.$$

By examining the impact of design-induced missingness on our ability to impute $\mathbf{E}R_t$ and the degree distribution, we obtain a sense of how information on network structure is lost as we look farther into the past.

IntervalN Design

The results from previous sections apply here as well. Querying the network cross-sections within the measurement interval gives us perfect knowledge of both edges and nulls, and thus nothing need to be imputed. On the other hand, if querying outside of the measurement interval, imputation on such network cross-sections fails to yield any useful result. This is because 1) such network cross-sections have very few observed edges and no observed nulls, providing us almost empty canvases to perform imputation on, and 2) with this much missingness, we have little information to build a reliable model. We can hence trivially characterize the performance of the intervalN design on imputing network properties as either contributing no error (for queries within the measurement interval) or making imputation

essentially impossible (for queries beyond the measurement interval).

LastK Design

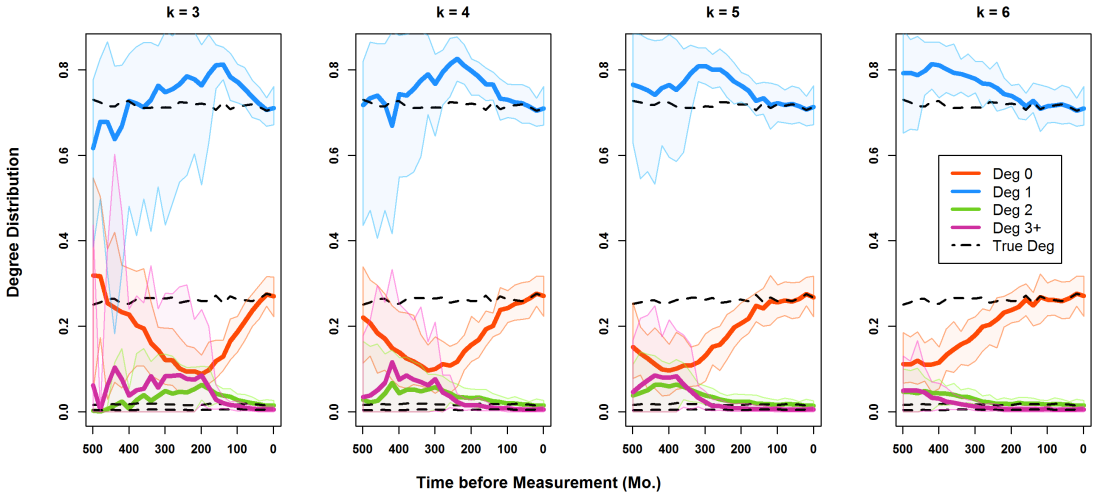

Figure 2.10: Imputed degree distribution as a function of look-back time under lastK designs. Vertical axis shows fraction of vertices having the specifed degree; horizontal axis shows look-back time. Mean imputed distribution values are shown in solid lines, with shaded areas depicting 95% simulation intervals. Dotted lines show true values.

Sexual contact networks within the general population show high degree 1 prevalence, followed by degree 0 (e.g., [Laumann et al., 1996, Hubert and Bajos, 1998, Youm and Laumann, 2002]). The dotted lines in fig. 2.10 illustrate this expectation: 72.9% of nodes have degree 1, 25.5% have degree 0, and the remaining 1.7% are concurrent. When a network cross-section is close to the time of measurement, the imputed degrees stay very closely to the true degree. As we move further away from time of measurement, the differences grow in such way that out imputation underestimate degree 0 and overestimate the higher degrees. This holds for all K's; it is also very clear that when K gets larger the imputed degrees remain accurate at longer look-back times.

As shown in fig. 2.11, the expected number of instantaneously reachable vertices ($\mathbf{E}R_t$) of the true network is relatively stable with small fluctuations around value 1.84. $\mathbf{E}R_t$ is fairly

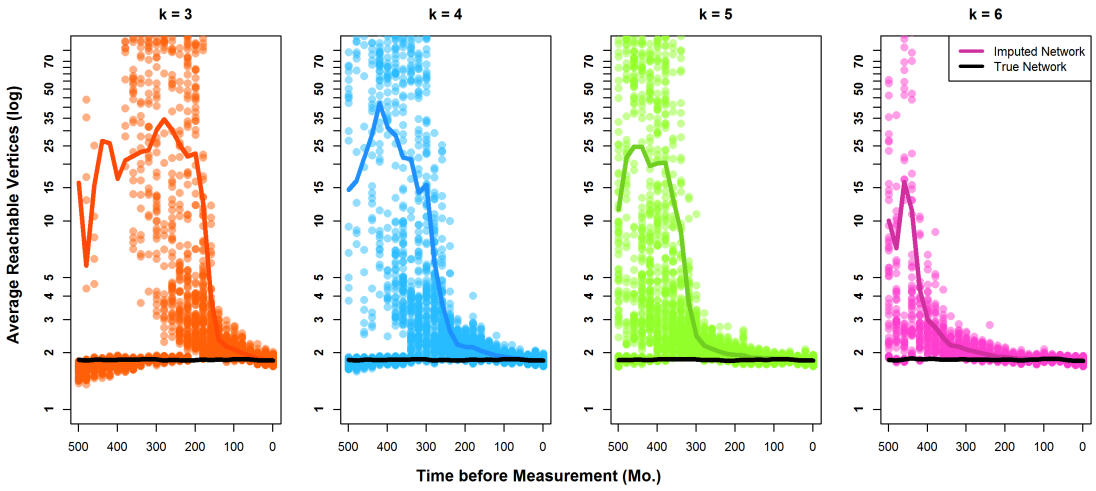

Figure 2.11: Imputed average instantaneously reachable vertices by look-back time. Dots are the imputed average reachable vertices from simulation draws; solid lines show the mean across all imputed networks. The average reachable vertices of the true networks are also plotted as references.

accurately imputed for short to moderate look-back times, with sudden collapse occurring at a point in time that scales with K. Beyond this point, the reachable vertex count is often overestimated by a wide margin (a side effect of concurrency overestimation). Since $\mathbf{E}R_t$ is closely related to epidemic potential, such an overestimate could in a real scenario lead to extremely misleading conclusions regarding the past history of the network.

2.5 Discussion and Conclusions

In this chapter we have characterized two basic classes of retrospective life history designs–intervalN, where subjects are asked to provide life events within a time interval of length N prior to the measurement time (e.g. time of interview), and lastK, where subjects are asked to provide last K life events–and examined their impact on retrospective network inference. We also examined simple variants of both designs (e.g., onset versus terminal selection). Analysis of what is and is not captured by these designs led immediately to a number of highly general insights, such as the differential capture of edges and nulls in each, the sharp

versus gradual failure modes of intervalN versus lastK as a function of look-back time, and the coverage advantages of terminal versus onset selection for lastK designs. These insights are not particular to the type of network being studied, and indeed may also be useful for understanding the consequences of RLH designs for spell data in a non-interpersonal context (e.g., employment or educational spells).

To gain additional insight into how these designs impact the loss of information as one queries farther into the past, we performed a simulation study based on sexual contact network data from the NHSLS. By using synthetic ground-truth data, we are able to isolate the pure impact of design missingness from other factors that may degrade performance in the real world, giving us a clearer sense of the inherent limits on inference posed by the designs. Information loss was considered by reference to three different types of outcomes: 1) simple accumulation of missingness; 2) impacts of missingness on parametric inference; and 3) impacts of missingness on retrospective network imputation. These three outcomes trace a pathway from the immediate loss of data to the impact of this data on substantive conclusions. While raw accumulation of missing edges and nulls is the most straightforward metric to evaluate designs, this does not directly convey the extent of information loss when modeling techniques are used. Error in inferred ERGM parameters for models fit to retrospectively queried cross-sections is a more abstract measure of information content, but conveys the extent to which missingness distorts the patterns present in the original data. Lastly, retrospective imputation allows us to assess how much of our knowledge of past network structure has been lost due to the impact of the design, giving us a direct measure of the limits to our ability to peer into the past when using RLH designs.

Our results have shown that the impact of intervalN design is almost binary: For query times within the measurement interval, no missingness is introduced, and hence there is no inherent impact of the design on retrospective inference or imputation. Outside this interval, however, little or no information is obtained. In the (left) censored variant, all edges outside

the measurement interval are truncated, thus no spell information may be obtained. In the non-censored case, although all nulls are are missing outside of the measurement interval, we do observe a small number of edges that are those who overlap with the measurement interval. Nonetheless, such observed edges are too sparse to be useful, and without information on nulls it is effectively impossible to estimate network density.[2] We hence regard retrospective inference to be effectively prohibitive for look-back times longer than the measurement interval. Since this property stems directly from the inherent missing pattern of the design (and not network structure), this conclusion is not SCN specific; indeed, it will apply to any network.

In the case of lastK designs, information loss is much more gradual. We nevertheless find that missingness among nulls is unacceptably high for onset selection, limiting its utility. In the appendix (sec. A) we show that terminal selection is almost always superior subdesign when compared with onset selection, thus our modeling and imputation focused on terminal section subdesigns.

For terminal-selected lastK designs, our simulation study paints a consistent picture of decline in information as a function of look-back time. For queries close to the measurement time, very little information appears to be lost, and retrospective inference (either in terms of model parameters or imputation) is quite accurate. As one moves beyond a "safe zone" near the measurement time, missingness accumulates at an increasing rate, and retrospective inference begins to degrade. This degradation is often sharp, and is marked by considerable instability (with some samples yielding reasonable results, while others yield results that depart markedly from those based on the complete data). The size of this "safe zone" (and the speed at which results subsequently degrade) is dependent upon K, with higher K levels yielding good performance at substantially longer look-back times. As with the behavior of

[2]In an ERGM setting, the MLE for the edge term (which sets the baseline expected density) will not exist; while placing a prior on graph density could in theory allow Bayesian inference, the likelihood provides so little information that this does not appear to be a practical option.

intervalN designs, these qualitative properties stem from basic aspects of what the design does or does not select, and are expected to hold for any network. However, the quantitative question of how far back one can go before leaving the "safe zone" depends on network structure, and may vary in practice. When in doubt, performing pilot simulation studies such as those performed here before selecting a choice of K may be a wise step when designing an RLH instrument for use in a specific case.

While our simulation study focused on sexual contact networks as a natural and substantively important case, the above comments should make clear that many of our are by no means limited to SCN data. In fact, any retrospectively collected spell data with by-design missingness caused by limiting the number of spells recorded, or the timing of spells, is subject to having the missingness patterns and impacts discussed in this chapter. This phenomenon does not only apply to interview/survey collected spell data; for instance data queried from online social networks are not immune due to possible data restrictions each OSN platform have placed on the data collector.

Although by-design missingness in many settings is inevitable, studies under different circumstances may have different focuses that could suffer less from one design or another. We offer some practical recommendations on the choice of retrospective designs to collect network data in SCNs or other, similar, contexts:

IntervalN designs are well-suited to studies with a narrow temporal focus. Since they generate no design missingness during the measurement interval, intervalN designs are optimal when seeking to gain information on relationships and/or network evolution within a specified interval. Although it would seem that many studies could benefit from using this design to guarantee full information, the measurement interval must be set in advance and narrow enough that elicitation is possible in practice. For instance, studies have shown that the degree distributions for many spell datasets such as friendship or sexual

partnership network are highly skewed [Handcock and Jones, 2004, Mislove et al., 2007] and data collection instruments (e.g. survey or interview) must be constructed so as to aid recall for respondents with many spells (e.g., the survey instrument must create enough spaces for respondents to list all their spells). The obvious trade-off with intervalN designs, aside from the cost of maintaining perfect information from an interval, is that they yield little or no information on network structure before the measurement period, limiting the uses to which the data can be put. If the initially selected interval is discovered to be too narrow to meet the researcher's requirements (or if such requirements later change), the resulting data may not be usable for its intended purpose.

LastK designs can work well for open-ended investigations, but look-back ability depends on the nature of the spell data, and K should be kept large. LastK designs impose different *full information* intervals for every subject (as compared to intervalN designs, which create a global full information interval). Although they generate missingness at earlier times, information obtained from them degrades more gradually. Since such degradation is driven by relationship turnover, the method is to an extent "self-tuning" for the timescale of network dynamics (which is helpful if this is not known *ex ante*). While it is obvious that the average duration of spells plays an important role in how long the safe look-back zone is, a more important factor is such spells' tendency to co-occur (which is also closely related to the degree distribution of the network). For example, when using terminal selection, the K-th most recent spell's terminus marks the start of the perfect information interval (i.e., we have full information from this point to the measurement time). We show that applying network modeling techniques has the ability to remedy some of the missingness problems, and this ability depends on the length of spells. In general, it is important to select K to be large enough that few respondents will have the onset of their $K-1$th relational spell (at which point missing edges are introduced) too close to the measurement time. While information loss under lastK is more gradual than intervalN, it should also be

borne in mind that inferential quality can fall sharply and somewhat unpredictably at longer look-back times; examining the level of design missingness prior to querying may provide a rough guide to how far back one can go in any given case.

As a final note, we reiterate that the present study has focused entirely on the consequences of missingness arising from RLH designs themselves. Real-world studies will typically feature egocentric sampling (hence obtaining data from a limited subset of individuals), and information will also be lost (or corrupted) due to failures of memory or other sources of reporting error. A natural question for further research is how these other sources of error or missingness interact with the design effects studied here; it is natural to suspect, for instance, that false negative rates will increase with look-back time, which may ironically extend the temporal reach of lastK designs (because the last Kth recalled spell may in fact be the $K' > K$th actual spell in ego's history) while simultaneously adding measurement error. Similar complexities have been found in e.g. complete ego net designs [Almquist, 2012], and characterizing them requires not only a good baseline simulation of network structure, but also a good model for reporting error. Peering into the past is a difficult challenge under the best of circumstances, and there are limits to what can be recovered from retrospective interviews. Nevertheless, appropriate choice of RLH design can make the most of what information exists to be recovered, and may in some circumstances permit retrospective inference over fairly long periods prior to the point of measurement.

Chapter 3

Local Graph Stability in Exponential Family Random Graph Models

3.1 Introduction

One motivation for complex network models has been the elucidation of the connection between local and global aspects of network structure. For instance, the frequency distribution of triadic subgraphs strongly constrains higher-order structures like ranked clusters [Holland and Leinhardt, 1971], and partnership concurrency is closely related to forward connectivity in time-varying networks [Morris et al., 2009]. Biases in subgraph frequencies are themselves directly related to the conditions under which the state of one edge depends upon another [Frank and Strauss, 1986, Pattison and Robins, 2002, Snijders et al., 2006], creating a direct link between local processes that e.g. favor or inhibit tie formation or triadic closure and higher-order structure. The exponential family random graph modeling (ERGM) framework (discussed in detail below) has become a widely used approach for identifying and exploring such connections between local and global structure in social and other networks [Robins

et al., 2005, Lusher et al., 2012]. While it can be seen simply as a flexible language for specifying distributions on graph sets, ERGMs can also be interpreted as parameterizing a set of biases influencing relational structure, with realized networks emerging from the interplay of these biases; these biases are formally analogous to forces in a physical context, an analogy that has been exploited in applications of ERGMs to biophysical systems (e.g. [Grazioli et al., 2019b,a]). In some cases, these biases can also be interpreted in terms of utility theory [Snijders, 2001], with network structure arising from the equilibrium of a latent stochastic choice process in which agents' decisions to add or remove ties are shaped by the associated biases. One potential use of ERGMs is hence to probe the conditions that are sufficient for the emergence or persistence of particular types of network structure, particularly where multiple mechanisms may be simultaneously at work.

In this paper, we examine one facet of this latter question, specifically introducing a basic notion of local *network stability* vis a vis an ERGM family and characterizing the subspace of parameters for an arbitrary family that renders a target structure stable with respect to a set of alternative networks. As we show, the stabilizing region of the parameter space forms the interior of a convex cone originating at the origin, whose faces are associated with a subset of the alternative networks against which the target is being compared. These results are presented in section 3.2, along with a practical algorithm for efficiently finding the stabilizing region of the parameter space. Our stability analysis is then illustrated with a simple and intuitive example involving an idealized centralized group structure (section 3.3)), followed by an application to a well-known study of collaboration within a legal firm (in section 3.4). In both cases, the correspondence between our notion of local stability and persistence of structures under Metropolis dynamics (widely employed in Markov Chain Monte Carlo simulations of network structure) is explored, and the use of stability calculations to predict likely or unlikely edge changes is demonstrated.

Throught this chapter, we denote vectors by lower case boldface letters, as in $\mathbf{t}$, matrices by

uppercase boldface letter, as in $\mathbf{M}$. $|S|$ is cardinality of a set S.

3.2 Stability

Here we introduce a simple notion of local stability for network structures. The intuition is as follows. Assume we have a graph whose stability is to be assessed (the *target network* or *target graph*) with respect to a set of *alternative graphs* that might be observed, as well as a model family whose elements are probability distributions on a graph set that includes the target graph and the alternative set. Our goal is to find a subset of models within the model family under which the target graph is more probable than any of the graphs in the alternative set. When this subset is non-empty, its members are said to *stabilize* the target graph vis a vis the alternative set, and by turns the target graph is said to be *locally stable* vis a vis the alternative set under any model within the subset.

It should be noted that our approach is very general: the target graph is given to us from the outset; there is no limitation on the choice of the alternative set; and although parts of our discussion require the model to have some specific functional properties, it is not limited to only the exponential family models. However, we model families in ERGM form admit a particularly elegant characterization of stability, and we employ this framework throughout this paper.

In the remainder of this section, we first formally define our notion of stability, and show how the use of the ERGM formalism allows the stabilizing subset of models within a given family to be easily characterized. We discuss some important properties of this stabilizing subset (including its geometric representation), and then provide an efficient approach to computing it in practice.

3.2.1 Definitions

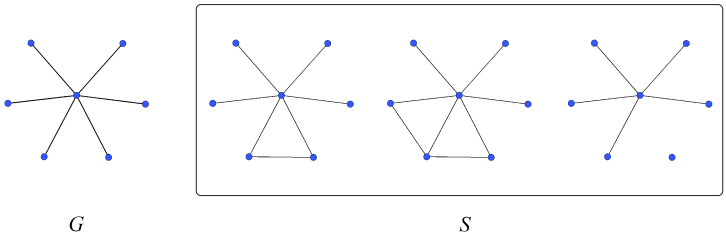

Figure 3.1: An example of a target graph G and alternative set S.

Let $\mathcal{G}$ be a finite set of graphs, $G \in \mathcal{G}$ be a *target graph* whose stability is to be assessed, and $S \subseteq \mathcal{G} \setminus G$ an *alternative set* of graphs with respect to which G is intended to be stable. For clarity of illustration we will emphasize the case in which all graphs in $G \cup S$ share the same vertex set, although this is not assumed. Likewise, we illustrate our ideas on simple graphs, but our development applies equally to directed, weighted, or multiplex networks.

We assess stability with respect to a model family on $\mathcal{G}$, which we take to be specified in ERGM form:

$$\Pr(G = g|\boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta} \cdot \mathbf{t}(g)\right)}{\mathcal{K}(\boldsymbol{\theta})}, \text{ where } \mathcal{K}(\boldsymbol{\theta}) = \sum_{g \in \mathcal{G}} e^{\boldsymbol{\theta}^T \mathbf{t}(g)}, \tag{3.1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^K$ is a vector of parameters and $\mathbf{t} : \mathcal{G} \mapsto \mathbb{R}^K$ is a vector of sufficient statistics. Intuitively, each element of $\mathbf{t}$ corresponds to a graph property whose distribution in G is to be biased, with the corresponding element of $\boldsymbol{\theta}$ indicating the strength and direction of the bias in question; in particular, $\mathbf{E}_{\boldsymbol{\theta}} t_i(G)$ is monotone in $\boldsymbol{\theta}_i$. Our goal is then to find the set of coefficients $\Phi = \{\boldsymbol{\theta}\}$ that make G more probable than any of the graphs in S, i.e.

$$\boldsymbol{\theta}, \text{ such that } \Pr(G|\boldsymbol{\theta}) > \Pr(G'|\boldsymbol{\theta}), \text{ for } G' \in S \tag{3.2}$$

with $\Phi = \{\varnothing\}$ if no $\boldsymbol{\theta}$ satisfies this condition. Any $\boldsymbol{\theta} \in \Phi$ is said to *stabilize* G with respect

to S, and Φ defines the *stabilizing subfamily* of the model family parameterized by $\mathbf{t}$ for G with respect to S. For convenience, we also refer to Φ as the *stabilizing subspace* of the full parameter space, noting the equivalence of models and parameter vectors in this representation.

Characterization of the Stable Subspace

Using (3.1), we can rewrite the inequality in (3.2) as

$$\frac{e^{\boldsymbol{\theta}^T \mathbf{t}(G)}}{\mathcal{K}(\boldsymbol{\theta})} > \frac{e^{\boldsymbol{\theta}^T \mathbf{t}(G')}}{\mathcal{K}(\boldsymbol{\theta})}$$

allowing us to simplify the stability condition as

$$\boldsymbol{\theta} : \boldsymbol{\theta}^T(\mathbf{t}(G') - \mathbf{t}(G)) < \mathbf{0} \text{ for } G' \in S. \tag{3.3}$$

The quantity $\boldsymbol{\theta}^T \mathbf{t}(G)$ is called the ERGM *potential*, and is equal to the log probability of G up to an additive constant. $\boldsymbol{\Delta}(G, G') = \mathbf{t}(G') - \mathbf{t}(G)$ is called the *change score*, and describes the way in which the sufficient statistics differ between graphs. We may then construct a matrix $\mathbf{M}$ by accumulating the change score vectors for G versus all G' in S. Specifically, we define $\mathbf{M}$ to be a $|S| \times K$ matrix where the i-th row is the change score $\Delta(G, G'_i)$, and the j-th column is the change scores regarding the j-th sufficient statistics:

$$M_{ij} = \Delta_j(G, G'_i), G'_i \in S.$$

With this notation it becomes clear that we can easily describe the local stability problem

algebraically as

$$\begin{aligned} &\text{Find all } \boldsymbol{\theta} \\ &\text{Such that } \mathbf{M}\boldsymbol{\theta} = \mathbf{v} \in \mathbb{R}_{-}^{|S|} \end{aligned} \tag{3.4}$$

with the set of $\boldsymbol{\theta} \in \mathbb{R}^K$ that satisfy the above constraint defined as Φ.

The algebraic characterization of Φ immediately reveals several useful properties of the stabilizing subspace. Trivially, the matrix product $\mathbf{M}\boldsymbol{\theta}$ in Eq. 3.4 can be rewritten as a set of row-wise inner products arising from members of the comparison set:

$$\sum_j M_{ij}\theta_j < 0, \text{ for } 1 < i < |S|.$$

Each corresponding inequality represents an *open half space*, whose *dividing hyperplane* (eg. $\sum_j M_{ij}\theta_j = 0$) passes through the origin. Intuitively, each half-space represents the portion of the parameter space under which G is more probable than a particular member of S. The intersection of these open half-spaces, if nonempty, is the interior of a convex polytope cone emanating from the origin (Fig. 3.2). We call this convex cone the *stable cone* of G w.r.t the alternative set S, since any parameter vector within it stabilizes G.

We can also prove the set Φ is a convex cone by showing that if $\theta_1, \theta_2 \in \Phi$ then $\alpha\theta_1 + \beta\theta_2 \in \Phi$, for any α and $\beta > 0$. *Proof*: Let $v_1 = M\theta_1$ and $v_2 = M\theta_2$, then $v_1, v_2 \in \mathbb{R}_{-}^{|S|}$. So,

$$\begin{aligned} v' &= M[\alpha\theta_1 + \beta\theta_2] \\ &= \alpha v_1 + \beta v_2 && \text{because } \alpha, \beta > 0 \\ &\in \mathbb{R}_{-}^{|S|} \end{aligned}$$

□

A practical implication of this observation is that the stabilizing subspace can be character-

ized in terms of a set of vectors (the directions of the facial intersections of the stable cone), from which any member of Φ can be obtained. In practice, we might expect that redundancies in the constraints implied by M will limit the number of vectors that are needed, an idea that we exploit below.

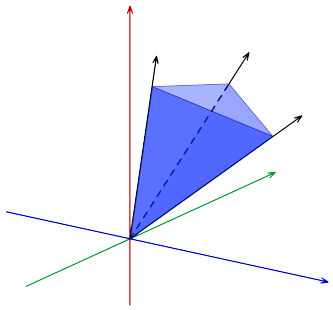

Figure 3.2: An example of the convex cone formed by three hyperplanes, in three-dimensional space.

3.2.2 Local Stability

The above stability definition can be applied to any arbitrary set of alternative graphs up to and including $S = \mathcal{G} \setminus G$; in the latter case, stability of G under $\boldsymbol{\theta}$ is obviously synonymous with G being the mode of the model specified by $\boldsymbol{\theta}$. Typically, however, we are interested in evaluating the stability of G with respect to a set of alternatives that are "close" to G under a hypothetical change process (e.g., the stochastic choice process of Snijders [Snijders, 2001], the reaction kinetics of Grazioli et al. [Grazioli et al., 2019b], or the Metropolis dynamics frequently used in MCMC algorithms [Hunter et al., 2012]) in which the network of interest evolves via discrete changes in which single edges are added or removed from the graph (*dyad toggles*) such that "uphill moves" on the probability surface occur at a higher rate than "downhill moves." Here, we discuss a specification of S that is broadly useful for assessing local stability in such settings. The intuition is as follows. Let S be the set of all graphs in $\mathcal{G}$ reachable from G by a single dyad toggle (i.e., the Hamming sphere of radius 1 centered on G). When G is stable with respect to S, moves away from G will be disfavored

(and return moves will be at least somewhat favorable); by contrast, when G is unstable with respect to S, there will be favorable dyad toggles that move away from G (with the return move being unfavorable). This notion of local stability (equivalent to G being a local mode in the Hamming space on $\mathcal{G}$) generalizes naturally to higher Hamming radii, and is a natural and easily computed starting point for considering longer trajectories (see fig.3.3 and fig.3.4).

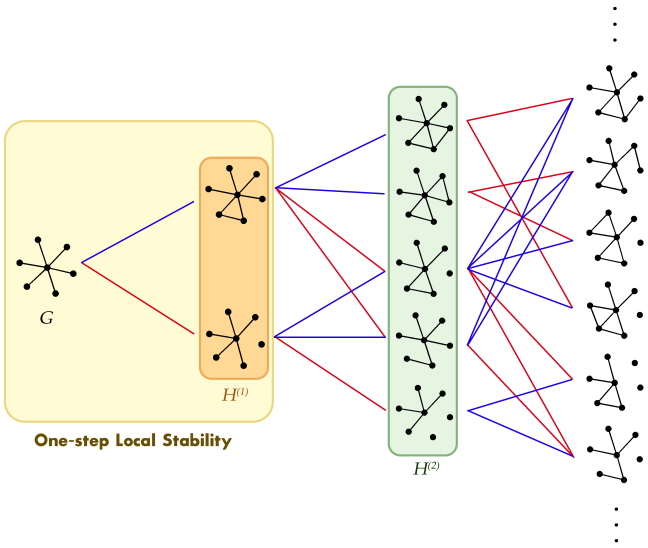

Figure 3.3: Examples of Hamming trajectories that could be taken from a target graph G. Here we group all isomorphic graphs into the same representation, and we only consider trajectories without loops. Each step is shown as either a blue line or a red line. Blue lines indicate steps involving edge addition and red lines indicate edge deletion. In this example, the set $H^{(1)}$ only contains two graphs, and local stability is satisfied if moves to either of the two graphs are disfavored.

Specifically, let $H^{(d)}$ be a set of all graphs that are Hamming distance d away from the target graph G. G may be said to be locally stable at (Hamming) radius d if G is stable with respect to $S = H^{(d)}$. In the special case of a single edge change, we are interested in $d = 1$. Since the number of rows in M is the cardinality of S, the stable cone for a simple graph with v vertices arises from the intersection of $\frac{1}{2}v(v-1)$ half-spaces. Although this quadratic scaling (more generally, $\mathcal{O}(v^{2d})$) is unfavorable, it is typically the case that symmetries associated with $\mathbf{t}$ lead many rows of M to be identical; such redundant rows can be removed without changing the solution to the stability problem. In some cases, the resulting compression can

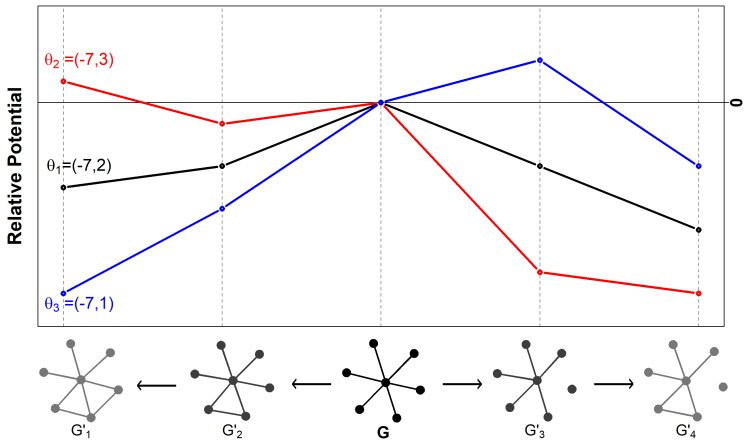

Figure 3.4: Illustration of a local mode in Hamming space. We choose two example trajectories, both started from the target graph G (plotted in the center). Trajectory G-G'_2-G'_1 is plotted from G to left and trajectory G-G'_3-G'_4 is plotted from G to right. Relative potential is the potential gain moving away from G. The model here has two terms: `edges` and `kstar(2)`. We choose three different θ's, and the relative potentials for each choice of θ are plotted in different colors. Under the model with coefficients θ_1 (in black), G has highest potential among these five graphs. In fact, G has the highest potential among all graphs in $H^{(1)}$ and $H^{(2)}$, indicating that G is stable against any two-step toggles. Under θ_2 (in red), G is *locally stable* because it has higher potential than both graphs in $H^{(1)}$. Note that although G'_1 has higher potential than G under the coefficients θ_2, trajectory to G'_1 is separated by graph G'_2 whose potential is lower than G. Under the coefficients θ_3 (in blue), G is considered unstable, due to the fact move to G'_3 increase the potential. In fact, coefficient θ_3 stabilizes G'_3 instead.

be considerable: for instance, one of the examples we show in section 3.3.1 leads to a two-row M matrix regardless of v. Moreover, some unique rows of M may also be redundant, in that they specify constraints on the stable region that are weaker than the constraints imposed by the other rows of M. As we show below, this form of redundancy can be exploited to calculate the stable cone in practical settings (as illustrated in section 3.4).

3.2.3 Solving for the Stable Cone

We employ the following terminology to refer to geometric features associated with the stability problem. Assuming we are in a K-dimensional parameter space, a $K-1$-dimensional dividing hyperplane separates the region of the parameter space where the target graph is

more probable than a specific alternative graph $G' \in S$ from the region of the parameter space where the target graph is at most equiprobable to G'. A dividing hyperplane may or may not constrain the stable cone (the intersection of half-spaces imposed by all dividing hyperplanes associated with the rows of M). When it does, the "side" of the cone created by its intersection with the hyperplane is called a *facet*. An intersection of two facets is called a *ridge*, which is an element of dimension $K - 2$.

The stable cone can be characterized using two different geometric representations: the *halfspace representation* (or *H-representation*) and the *vertex representation* (or *V-representation*) [Avis and Fukuda, 1992]. The H-representation characterizes the subspace Φ as a set of conditional linear inequalities, or half-spaces, as shown by the blue facets in Fig. 3.2, and in Eq. 3.5 below:

$$\Phi = (\theta | M_i \theta < 0) \ \forall i \in \{1, 2, ..., k\} \tag{3.5}$$

while the V-representation characterizes the subspace as the convex hull generated by the vertices that are created by each pair of intersecting planes from the H-representation (shown as black arrows in Fig. 3.2). It is worth noting that, while solving for the stable region, it is convenient to store each vertex as a single point in parameter space by calculating the intersection of each vertex with a hypersphere of a given radius (shown in Fig. 3.5). Since all points along the ray associated with each vertex can be obtained by rescaling, as can the ray associated with any point within this particular slice of the stable region (solid orange triangle in 3.5), no information is lost by using this representation. Additionally, the data structure for the V-representation is such that the indices of the hyperplanes whose intersection comprise each vertex is also stored in the V-representation object. The remainder of this section presents the calculation of the stable region using this method for storing the V-representation.

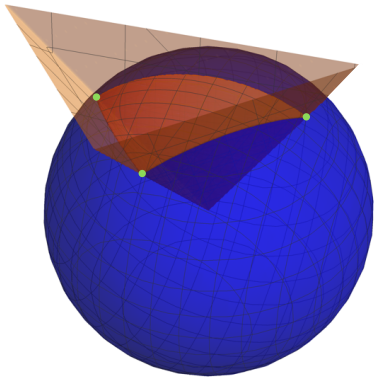

Figure 3.5: The vertex representation (intersections of the transparent orange planes) can be conveniently stored as the points (green dots) produced by the intersection of the vertices with a hypersphere (blue sphere)

Although the H and V representations each offer a complete description of the stable cone individually, they are perhaps most useful for solving the stable cone when the two representations are used in tandem. In our application, we leverage the complementarity of the two representations to iteratively search for both a non-redundant H-representation and V-representation simultaneously by applying a methodology based on the *double description method* (or *DD method*), introduced by Avis [Avis and Fukuda, 1992]. Given an initial redundant $\mathbf{M}$ matrix (*redundantHalfspaces*), with all edge changes of interest represented in the rows, the stable cone for a given target graph and set of sufficient statistics can be solved by first using Algorithm 1 (Fig. 3.6) to obtain an initial closed superset of the stable region, which is then passed (along with $\mathbf{M}$) to Algorithm 2 (Fig. 3.7), where the initial closed superset is whittled down to a non-redundant double description of the stable region.

Given the redundant $\mathbf{M}$ matrix for a target graph and a set of sufficient statistics, this information is first passed to Algorithm 1 in order to calculate an initial closed superset of the stable region. Algorithm 1 begins by first obtaining a set of $|\theta| - 1$ rows from $\mathbf{M}$, where no two halfspaces are parallel, and calculating their intersection with both each other and the hypersphere. These rows and this vertex point serve to initialize the H-representation (H) and V-representation (V), respectively. At this point, a new row is drawn from $\mathbf{M}$ and appended to H, and intersections between all possible combinations of $|\theta| - 1$ rows of $\mathbf{M}$

are calculated and then appended to V. Although the current iteration of H is redundant, its stable region is still equivalent to the non-redundant form, thus H is used to remove all unstable vertices from V. Given that the data structure for V includes the labels for the rows of H that intersect to form each vertex in V, it is then trivial to now remove all rows of H that are not represented in V. At this point, H and V are both non-redundant. The final step in the loop is to check the convex hull for closure. If the convex hull represented by H and V is closed, Algorithm 1 returns H and V and terminates. One straightforward method for testing the closure of the convex hull described by V is to first check that the number of vertices is $\geq |\theta|$, and then use any of the many available methods for calculating convex hulls from points (e.g. Quickhull [Barber et al., 1996]) to calculate the convex hull of the vertices in V. If both the number of vertices is $\geq |\theta|$, and the number of halfspaces in the halfspace representation returned by the convex hull finding method is equivalent to the number of rows in H, the stable region represented by H and V is closed. This test for closure is made possible by the fact that if the convex hull finding algorithm is operating on a convex hull that is open with respect to the stable region, it will introduce a new halfspace to close off the open end of the space.

Algorithm 1: Finding an initial closed superset of the stable region

```
Data: M
initialize H; V;
bool are.parallel = TRUE;
while are.parallel do
    h.init = sample.two.rows(M);
    are.parallel = check.parallel(h.init)
end
H = h.init;
V = get.exhaustive.intersections(H);
bool hull.is.closed = FALSE;
while !hull.is.closed do
    h.test = sample.one.row(M);
    H = append(H, h.test);
    vertices.new = get.exhaustive.intersections(H, h.test);
    V = append(V, vertices.new);
    V = return.stable.vertices(V, H);
    H = get.H.from.V(V);
    hull.is.closed = testForClosedConvexHull(V)
end
return {H, V}
```

Once Algorithm 1 has returned an initial closed superset of the stable region, H, V, and $\mathbf{M}$ are then passed to Algorithm 2. The premise of Algorithm 2, as put forth by [Avis and Fukuda, 1992], is that given an initial closed convex hull, any newly introduced halfspaces are only non-redundant if their introduction excludes one ore more previously existing vertices. The algorithm also leverages the fact that the data structure for V keeps track of which

halfspaces intersect to produce each vertex, in that upon introduction of a new non-redundant halfspace, only the halfspaces whose intersections comprise the newly excluded vertex or vertices must be included in the calculation of newly created vertices. Intuition for the methodology can be readily obtained from Fig. 3.7.

Algorithm 2: Using the DD method to solve the stable cone

```
Data: H.initial, V.initial, M
initialize H = H.initial; V = V.initial;
for all rows in M do
    h.test = get.next.row(M);
    v.excluded = get.excluded.vertices(V, h.test);
    if is.not.empty(v.excluded) then
        H = append.row(H, h.test);
        rows.for.testing = get.rows.comprising.vertices(v.excluded);
        vertices.for.testing = get.new.intersections(rows.for.testing, h.test);
        v.new = get.stable.vertices(vertices.for.testing);
        V = append(V, v.new)
end
return {H, V}
```

The order of computational complexity for this methodology is best understood using a bounding argument. First, we establish the two fundamental operations employed by our methodology: 1) calculating intersections of sets of halfspaces that form vertices, and 2) testing whether or not vertices lie within the stable region. The first operation is an inversion of a $d \times d$ matrix, where d is the dimensionality of the model (i.e. the number of sufficient statistics), and the second operation is a simple matrix multiplication of a vector of length d by the matrix representing the current iteration of the H-representation (with trivial parallelization); thus, the rate-limiting operation is the calculation of vertices. Next we establish the scaling for a brute force treatment whereby, first, all n choose $d-1$ inter-

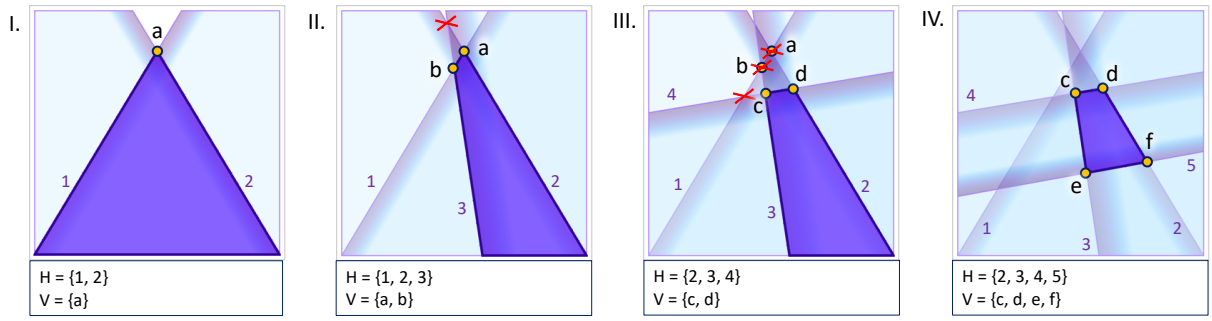

Figure 3.6: A demonstration of Algorithm I, used to define an initial closed superset of the stable region. I.) The first step in Algorithm I: the H-representation is initialized with two non-parallel halfspaces drawn from $\mathbf{M}$, while the V-representation is initialized as the intersection of the two halfspaces in H. II.) Step 2 in algorithm I: A third halfspace is introduced, all n choose 2 intersections are calculated, all intersections within the stable region defined by H become the V-representation, and all halfspaces whose intersections comprise V become H. The convex hull is not closed, so we iterate another step. III.) Step 3 in algorithm I: as in the previous step, a new halfspace is introduced, stable intersections become V, and their respective halfspaces become H. Note that in this case, two previously included vertices (a and b) are now excluded from V, as is halfspace 1, since it no longer contributes any stable intersections to V. Still, the convex hull is not closed, so we iterate another step. IV.) A new halfspace is introduced, and the process from the previous two steps is repeated. This time, the convex hull is closed, thus Algorithm 1 is terminated, and an initial closed superset of the stable is returned.

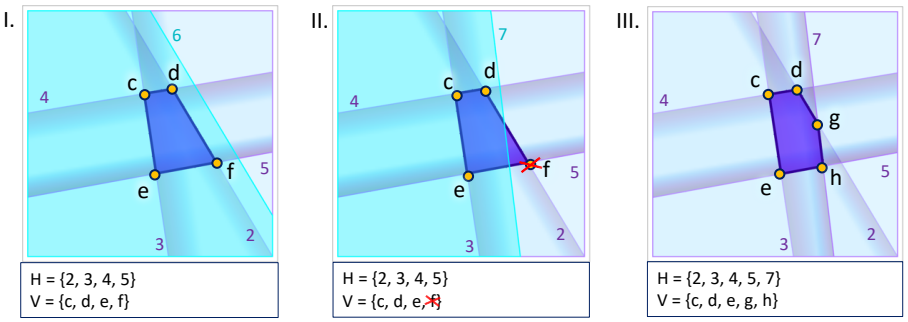

Figure 3.7: A demonstration of Algorithm II, which takes an initial closed superset of the stable region and an $\mathbf{M}$ matrix as input, and outputs H and V descriptions of the stable region for a given model. I.) The closed set from figure 3.6 (step IV) becomes the input for Algorithm II.) A new halfspace is introduced (cyan), but since no vertices from V are excluded, the halfspace is redundant, and thus rejected. II. Another halfspace is introduced, which excludes vertex f, thus f must be rejected from V and all intersections between the new halfspace and those whose intersections comprise f must be calculated and tested for stability under H. III.) The new halfspace is appended to H and the intersections between 2, 5, and 7, stable under H are appended to V.

sections between hyperplanes are taken to obtain all possible vertices (operation 1), then each vertex is tested to determine whether or not it is within the stable region (operation 2), leaving the non-redundant double description. In this case, the computational complexity scales as $\mathcal{O}((\frac{en}{d})^d)$. The worst-case scenario for our methodology would be that *Algorithm 1* fails to produce a closed stable region until the very last row from the M matrix has been included in the exhaustive search, demonstrating that, at its worst, the order of complexity for our algorithm is equal to the exhaustive brute force treatment. Next, consider the best-case scenario: given that a particular stable region can be expressed as an H-representation comprised of h rows, suppose that these are the first h rows selected from the M matrix in *Algorithm 1*. In this case, the only vertices calculated would be those produced by the H-representation implying a computational complexity of $\mathcal{O}((\frac{h}{d})^d)$. Even for a pessimistic case where only half of the rows in M are redundant ($h = nrow(M)/2$), our methodology would offer a computational speedup of $.5^{-d}$, a massive speedup over the fully exhaustive brute force method (e.g. $32\times$ speedup for a network model with just 5 sufficient statistics).

3.3 Case Study I: Cult (Star) Structure

In this section, we apply the above method to a simple example of a social structure that we here refer to as the *cult network*. A cult network is characterized by an isolated star structure, where there is a single core ("leader") node in the center connecting to all peripheral ("follower") nodes, and no peripheral nodes are mutually adjacent. Thus, a follower can reach any other follower, but only via a path that is brokered by the leader. Although the motivation behind introducing our methodology using a cult network as an example is illustrative simplicity, it does represent a stylized version of features found in some real-world charismatic cults. For example, both the People's Temple [Johnson, 1979] and Heaven's Gate [Davis, 2000] cults featured leaders who eventually isolated their groups from outside

contact and regulated the flow of critical information such that rank-and-file members were encouraged to trust and obey only the leaders themselves (with unmediated intra-group relationships among members being heavily discouraged). Maintaining these star-like group structures required considerable creativity and effort on the parts of the cult leaders, who ultimately crafted complex social environments that stabilized what would otherwise be highly unfavorable pattern of social relationships. The cult network model presented here demonstrates the conditions that are necessary for such a structure to be stabilized under one very simple class of social processes.

The model family employed here is by design minimal, incorporating only two types of social "forces:" a general propensity to form edges, and a force that governs the propensity for untied pairs of individuals to have or lack partners in common. In ERGM form, such a family corresponds to a vector of statistics ($\mathbf{t}$) containing respectively the count of edges (`edges`) and the count of null dyads having no partners in common (`nsp(0)`). We investigate whether the star structure can be stabilized under this model family, and if so, define the region of the parameter space where that the network is stable.

In the remainder of this section, we first compute the stable cone, then demonstrate that this region aligns with stability as assessed by simulated Metropolis dynamics. We also consider the question of which dyad is likely to be the first to be toggled (assuming that some toggle occurs). This approach can help identify which edges/nulls are most vulnerable to change under the model. We can also show analytically the probability of the target network transitioning to each of the alternative networks, given that an edge change has occurred. We then validate these results with simulated trajectories.

We use an undirected star network of v vertices as the target network. For illustrative purposes, we let $v = 7$ when figures or numeric results are generated (e.g., G in figure 3.8). For simplicity, we assume that no attribute is associated with vertices or edges, indicating that all peripheral nodes are interchangeable. Because our model is unable to differentiate

between the target graph and other members of its isomorphism class, all isomorphic graphs will be counted towards stability. The set $S = H^{(1)}$ contains only two distinct networks (illustrated in fig. 3.8): G_+ is the result of adding one edge to G, connecting two peripheral nodes; and G_- is the result of removing one edge from G, breaking one connection between the core and a peripheral node. As mentioned, the two sufficient statistics are `edges` and `nsp(0)`). We note in passing that there are other model families that can also generate star structures. For instance, a term influencing the number of null dyads having exactly one shared partner (`nsp(1)`) can also be used to generate star-like graphs. However, the `nsp(0)` term, despite being less obviously associated with the star structure, interacts with the `edges` term to anchor the structure in place, i.e. a negative value in `nsp(0)` suppresses the existence of non-edges sharing exactly zero partners, thus non-edge pairs with higher shared partners are boosted, to a limit that is controlled by the network density (set by `edges`).

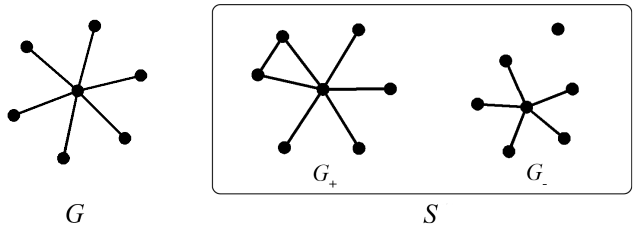

Figure 3.8: $S = H^{(1)}$ for the star structure.

3.3.1 Stable Parameter Region of the Star Structure

The rows of the M matrix are comprised of the change scores (i.e., differences of graph statistics) between the target graph and the graphs in the alternative set (shown in table 3.1). The parameter space of this model is 2-dimensional - each sufficient statistic of the model occupies one dimension. To illustrate the parameter space we plot `edges` on the horizontal dimension (x-axis) and `nsp(0)` on the vertical dimension (y-axis). For simplicity, hereafter

we use x, y to represent the parameter values of `edges` and `nsp(0)`, respectively.

	Sufficient Statistics (t)		Change Score (M)	
	`edges`	`nsp(0)`	`edges`	`nsp(0)`
G	v-1	0	-	-
G_+	v	0	1	0
G_-	v-2	v-1	-1	v-1

Table 3.1: Sufficient statistics and change scores for the star structure G and the set $S = H^{(1)}$. v is the number of vertices in the graph; `edges` and `nsp(0)` are the two ERGM terms used in the model.

As derived in section 3.2.1, each graph in S defines a half-plane (or more generally, a half-space; in subsequent discussions, we use general terms such as half-spaces and hyperplanes, despite them being called half-planes and lines in 2-dimensional space): $G+$ defines $x \leq 0$, and G_- defines $x-(v-1)y \geq 0$. The stable region, which is the intersection of the two half-spaces, is plotted as the shaded area in fig. 3.9a (assume $v = 7$). As proved in section 3.2.1, the stable region is a cone that has infinite height and points at the origin.

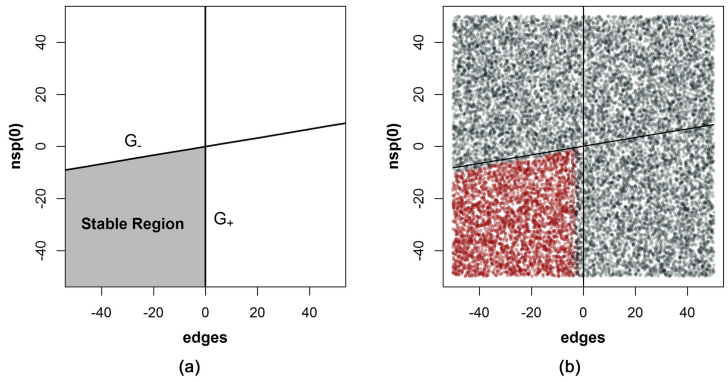

Figure 3.9: **a**. The stable cone (grey region) for the star structure under the `edges`/`nsp0` family, assuming $v = 7$. **b**. Fraction of simulated networks remaining in star configurations after 10^6 random-walk Metropolis steps for selected parameter values (dots). Color values range from grey (no persistence) to red (complete persistence). Despite being defined only relative to $H^{(1)}$, the stable cone (line-bounded region in lower left) closely matches the region of long-term dynamic stability.

To verify that the stable cone is compatible with stability under explicit dynamics, we sample 10^6 parameter vectors and run 50 random walk Markov chains at each sampled vector. (All simulation was performed using the `ergm` package within `statnet` [Hunter et al., 2008,

Handcock et al., 2008] using random dyad proposals.) Each chain starts from the perfect star structure and runs for $t = 10^6$ Metropolis steps. We examine whether the star structure is preserved in the networks returned at the end of the simulation, and plot the fraction of the networks that remain unchanged (fig. 3.9b).

Even though the stable cone is based only on stability versus $H^{(1)}$, it corresponds closely to long-term stability under Metropolis dynamics (red points in Fig. 3.9b). Outside of the stable cone, the graph is dynamically unstable (dark grey points in fig. 3.9b) Along the interior of the two facets, there is a thin stripe of models that are dynamically unstable despite being inside the locally stable region. This is reflective of the fact that, while moves away from G are unfavorable, they will eventually happen given enough attempts (here, 10^6). Parameter values close to the faces of the stable cone have a lower margin of stability, in the sense that probability gap between G and the elements of S is smaller, and dynamic stability hence begins to be lost as one moves from the center of the locally stable cone to its faces. An interesting observation is that the "unstable band" associated with G_+ is a bit wider compared with the band associated with G_-. This phenomenon can be explained by examining the ERGM potential of each alternative graph.

By definition, the stable cone is the region where the ERGM potential of the target graph, $\boldsymbol{\theta}^T \mathbf{t}(G)$, is greater than any of the other graphs in S; facets and ridges of the cone then correspond to the sets of $\boldsymbol{\theta}$ values where at least one or two graphs (respectively) in S have the same potential as the target graph. Because the graph probability is proportional to the exponentiated ERGM potential, $\Pr(g) \propto \exp(\boldsymbol{\theta}^T \mathbf{t}(g))$, examining graph potentials provides insight into the behavior of the Metropolis dynamics. As shown in fig. 3.10, the potential difference of G and $G_+$is much more gradual along the dividing hyperplane comparing with that of G and G_-, thus the down-potential step is taken with higher probability during the Markov Chain run.

The above simulation experiment considers whether network snapshots under a long simu-

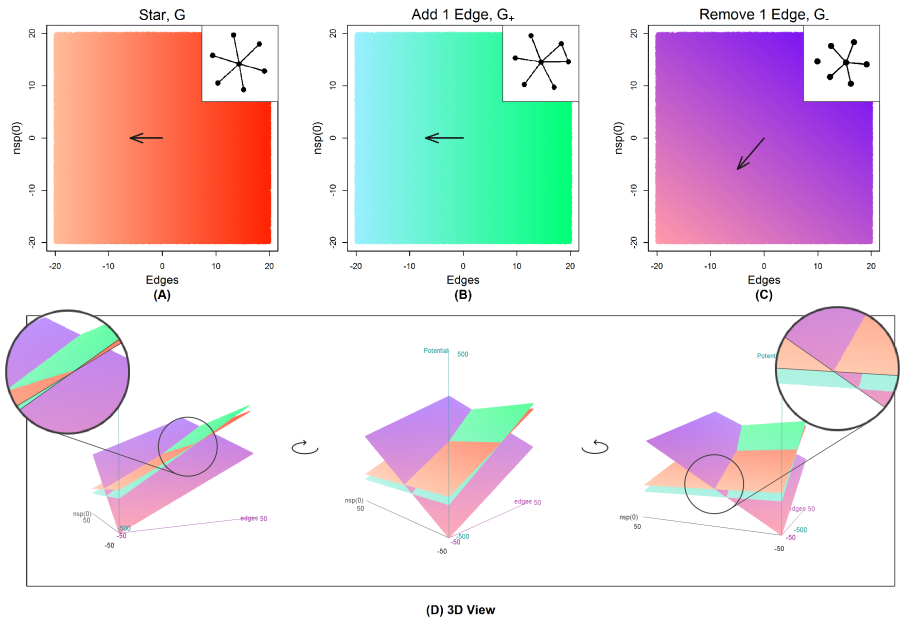

Figure 3.10: Potential planes of target graph G, and both alternative graphs G_+, G_- in S. Panels A, B, and C plot the potential of networks G, G_+ ,and G_-, respectively. The gradients of the potential surfaces are plotted as arrows. Panel D plots the three gradients onto the same space. Note that the stable cone can be identified in the bottom left quadrant, where the potential of G is higher than both G_+ and G_- (see inset).

lation are isomorphic to the target graph. In principle, a network could alternate between a number of states during the simulation, counting towards stability so long as it is in the target state when observed. A more rigorous test of dynamic stability is to examine the expected time to the *first* change under Metropolis dynamics across the stable cone. Fig. 3.11 shows the mean number of steps required for the first accepted toggle, from yellow (1 step) to red (> 1000 steps). As in the previous simulation, we see that the graph strongly resists perturbation within the bulk of the stable cone, staying unchanged for more than 1000 simulation steps. The wider stripe of dynamic instability along the $x = 0$ hyperplane ($P(G) = P(G_+)$) is associated with more rapid network changes, with fewer than ten simulation steps required on average as one nears the boundary of the stable cone. The relatively sharp transition from the less dynamically stable region to strong dynamic stability results from the exponential decline in acceptance probabilities with respect to the potential difference: while our definition of S leads to an extremely local definition of stability, the exponential decline in

transition probabilities as one moves away from the faces of the stable cone makes it a good proxy for dynamic stability more broadly defined.

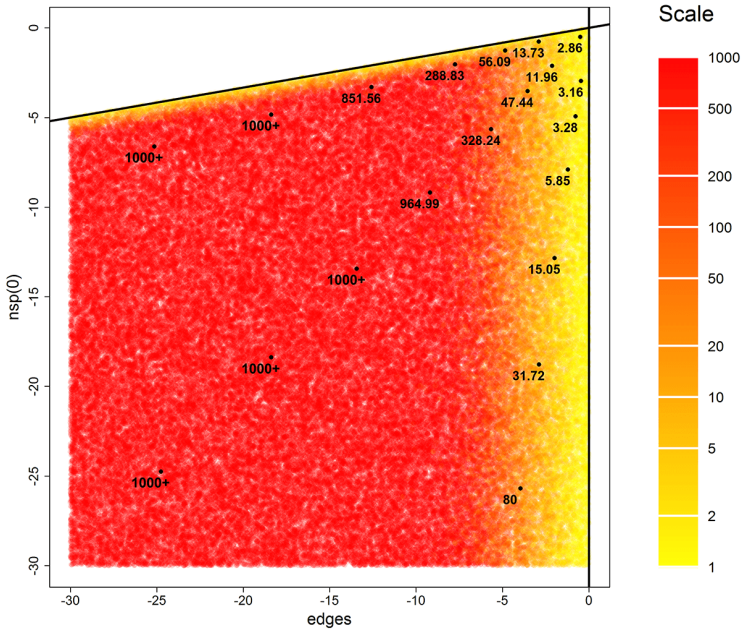

Figure 3.11: The average number of Metropolis steps required for the first change in G, as a function of $\boldsymbol{\theta}$. Waiting times grow exponentially fast as one moves away from the faces of the stable cone.

3.3.2 Dyad Vulnerability

We have shown that the target structure is most dynamically stable well within confines of the locally stable cone, with dynamic stability weakening in the region closer to the facets of the cone. Of course, given enough attempts, a change will inevitably occur; this raises the question of which structure in S will be selected when the first dyad toggle is accepted. This can be thought of as a problem of *dyad vulnerability*: if change happens, which dyads are most vulnerable to being toggled? In the case of the cult network, symmetry leaves us only two types of possible dyad toggles: toggles that break connections between the center vertex and one of the periphery vertices (denoted d_-, resulting in a graph isomorphic to G_-), or toggles that establish connections among periphery vertices (denoted d_+, resulting

in a graph isomorphic to G_+).

Assume a dyad toggle d_i is selected randomly from all possible $\frac{1}{2}v(v-1)$ toggles. The two types of toggles are proposed at rates based on the frequencies they appear in G: $\Pr(d_-) = \frac{2}{v}$, and $\Pr(d_+) = \frac{v-2}{v}$. Then the acceptance ratio ($\alpha = \Pr(G_i)/\Pr(G)$) is calculated. The proposed toggle is accepted if the new graph is more probable, i.e., $\alpha > 1$. Otherwise, the toggle is accepted with probability $\Pr(\text{accept}|d_i) = \alpha = \exp((t(G_i) - t(G)\theta^T)$. The acceptance probability if d_- is proposed is:

$$\Pr(\text{accept}|d_-, G) = \begin{cases} \exp(\boldsymbol{\theta}(t(G_-) - t(G))), & \text{if } \boldsymbol{\theta}(t(G_-) - t(G)) \leq 0 \\ 1, & \text{otherwise.} \end{cases}$$

Then the probability of the target graph G change to G_- can be calculated as

$$\Pr(G_-|G) = \Pr(d_-)\Pr(\text{accept}|d_-).$$

Similarly, $\Pr(G_+|G)$ can be derived when d_+ is proposed. When proposed moves are rejected, then graph stay unchanged. Let $\boldsymbol{\theta} = [x, y]^T$ where x is the parameter value for `edges` and y is the parameter value for `nsp(0)`, the probability of becoming G_+, G_-, or stay unchanged given starting from the perfect star structure G is:

$$\Pr(G_-|G) = \begin{cases} \frac{2}{v}e^{-x+(v-1)y}, & \text{if } x - (v-1)y \geq 0 \\ \frac{2}{v}, & \text{otherwise} \end{cases},$$

$$\Pr(G_+|G) = \begin{cases} \frac{v-2}{v}e^{x}, & \text{if } x \leq 0 \\ \frac{v-2}{v}, & \text{otherwise} \end{cases},$$

$$\Pr(G|G) = 1 - \Pr(G_+|G) - \Pr(G_-|G)$$

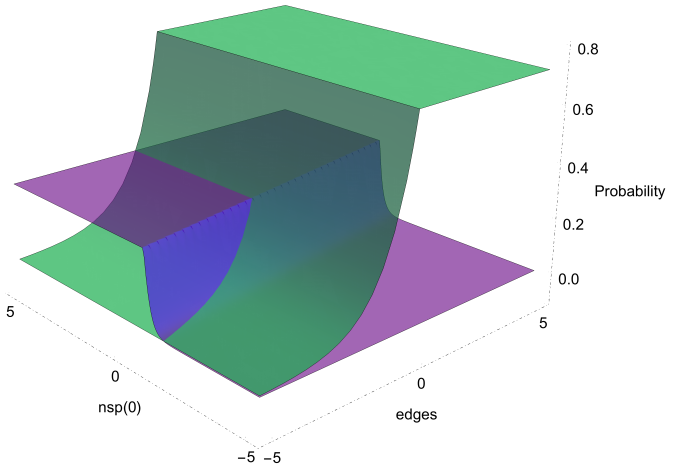

Figure 3.12: One step transition probability to each of the structures in $H^{(1)} = \{G_+, G_-\}$, as a function of the model parameters. The green surface corresponds to G_+, and the purple surface corresponds to G_-. The probability of staying unchanged as G is 1 minus the sum of the two surface values.

To illustrate the dyad vulnerability, we sampled 10,000 parameter vectors from $[-10, 10]$ interval for both `edges` and `nsp(0)` at $v = 7$. This region covers both the stable cone and the unstable region. We ran a 5000 step random walk MCMC trajectory for each sampled parameter until the first change occurred. The simulation result (in fig. 3.13) is in accordance with the theoretical derivation.

3.4 Lazega's Lawyer Dataset

In this section, we analyze network stability with a real-world dataset and a published model. We note at the outset that although perfect local stability (i.e., the stable cone w.r.t. $H^{(1)}$ is non-empty) is reached with the star example, not all model families lead to non-empty regions of local stability for $S = H^{(1)}$. For instance, consider a network with an attribute that classifies nodes to be in either group A or B, and a model family that has two terms, a general `edges` term to control the overall density and a `nodemix` term that counts the

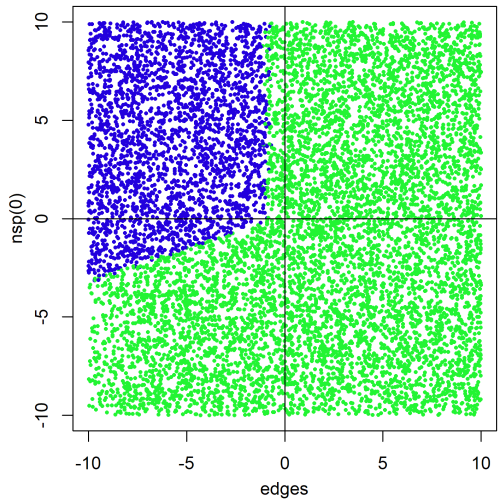

Figure 3.13: Simulation showing the most vulnerable dyad type, as a function of model parameters. Dots indicate sampled parameters, each of which was used to govern one MCMC trajectory; blue dots are networks that changed to G_- on the first toggle, indicating that breaking an existing connection is more likely, and green dots are networks that first changed to G_+, indicating establishing a connection between two periphery nodes are more likely.

number of between-group ties. If there exist vertices $(a, a') \subseteq A$ and $(b, b') \subseteq B$ such that edge $\{a, b\}$ is in G and edge $\{a', b'\}$ is not in G, then the change scores for toggling the $\{a, b\}$ and $\{a', b'\}$ dyads have respective elements `edges` = -1, `nodemix(A,B)` = -1 and `edges` = 1, `nodemix(A,B)` = 1. These two change score vectors define two half-spaces whose intersection is empty (i.e., $x + y > 0$ and $-x - y > 0$), and thus there is no model that stabilizes such a graph. Intuitively, this is because any choice of $\boldsymbol{\theta}$ for this family will on balance either favor adding or removing cross-group edges, and hence one of the two dyad states will not be favored. More generally, for any G and model family parameterized $\mathbf{t}$, if there exist an edge and a null in G with directly opposite change scores with respect to $\mathbf{t}$, then the local stable region associated with $S = H^{(1)}$ will be empty. In such cases, we can conclude that the forces governing the associated with $\mathbf{t}$ do not (or would not) locally stabilize G. Beyond this observation, we can gain additional insight into model behavior by examining the subsets of S for which stabilization *is* possible, particularly where $\boldsymbol{\theta}$ is known or has been estimated from the observed network [see e.g. Hunter et al., 2012]. Moreover, where the local stable cone is non-empty but an estimated model does not lie within it, we

can exploit the position of the estimated model relative to the stable cone to obtain insights into the factors that are driving instability, and the hypothetical changes in social forces that would lead the target graph to become locally stable. In this section, we illustrate how some of these techniques can be used to gain insights into model behavior in a non-trivial setting.

We study graph stability with a data set collected by Lazega [Lazega, 2001] on working relations among 36 partners in 1991 in a New England corporate law firm. This dataset is a network where edges (undirected) represent collaborations between partners. For purposes of analysis, we employ a model family for this data set that was previously published by [Hunter, 2007]. Due to improvements in estimation methods since the original publication, we here refit the model using the `ergm` package [Handcock et al., 2019] to obtain updated coefficients (table 3.2). The covariates used in the model are as follows: seniority, which describes the rank order of entry into the firm (1=earliest, 36=latest); type of practice (1=litigation, 2=corporate); the office at which the partner works (1=Boston; 2=Hartford; 3=Providence); and the partner's gender (1=man; 2=woman). The model includes the main effects of seniority and practice, along with homophily effects for practice, sex, and office location. A geometrically weighted edgewise shared partner (GWESP) term was also included to account for triadic closure.

Parameter	Estimate	S.E
Edges	-7.375	0.712^{***}
Main Seniority	0.024	0.007^{***}
Main Practice	0.411	0.118^{***}
Homophily Practice	0.761	0.192^{***}
Homophily Gender	0.696	0.256^{**}
Homophily Office	1.145	0.196^{***}
GWESP($\alpha = 0.75$)	0.937	0.159^{***}
$^{*}\ p < 0.05, \quad ^{**}\ p < 0.01, \quad ^{***}\ p < 0.001$		

Table 3.2: Maximum likelihood estimates for the Lazega model, with standard errors and Wald test p-values. α refers to the GWESP decay parameter, which was fixed to 0.75 during estimation.

We define $S = H^{(1)}$ (total 630 graphs), and compute the corresponding M matrix (630-by-

7). Each row in M is characterized by one graph in the alternative set, and each column is one statistic. The network is locally stable under the estimated model with parameter $\hat{\boldsymbol{\theta}}$ if $\mathbf{M}\hat{\boldsymbol{\theta}} \in \mathbb{R}_-$, i.e. if all 1-step changes are unfavorable. Interestingly, the observed graph is unstable against this alternative set, and the unstable fraction is 0.159 (100 out of 630 graphs). Because each graph in the alternative set corresponds to one dyad toggle and defines a stable half-space, then for each dyad there exists a stable half-space in which this dyad is more likely to stay unchanged. If a model lies within the stable half-space, then the dyad is said to be stable under this model, and vice versa. We calculate the distance from the model to stabilization of graph G_i' in $H^{(1)}$ as d_i (shown in fig. 3.14) and use the convention that positive d_i means the model lies outside of the stable half-space of graph G_i'; while negative d_i means the model is within the stable half-space of graph G_i'. The absolute value of d_i is the distance to the dividing hyperplane.

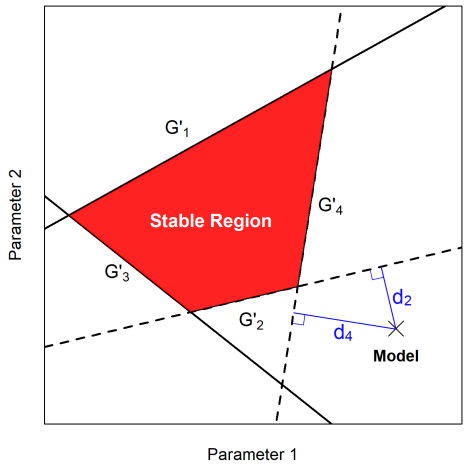

Figure 3.14: Example parameter space illustrating the stable region (in red), a particular model (black $\times$), and its distances to nearby hyperplanes (blue lines). Here, the target graph G is unstable under the model, as the model is positioned outside of the stable region. Nevertheless, the model is partially stable w.r.t. G_1' and G_3', because it lies on the side of the stable half-spaces of these two alternative graphs (bottom of G_1' and left of G_2'). In order to quantify the instability of the model, one can calculate the generalized distance from the model's position in the parameter space to each of the hyperplanes bounding the halfspaces for which the model is not included.

In fig. 3.15 we plot the stable edges, unstable edges, and unstable nulls. A few observations

can be made: 1) Two core groups are clearly identified, each centered within one of the larger offices (Hartford and Boston). Ties within groups are generally more stable (in blue) and ties between groups are mostly unstable (in red). 2) Both large offices demonstrate core-periphery structures internally. The cores (marked by shaded background) are characterized with stable connections, while the peripheries are connected by unstable connections. 3) There are relatively few unstable nulls, and most of them are within two dense groups. In fact, there is only one unstable null that connects two groups. 4) Within two offices, there are a decent amount of unstable nulls in the Boston office and only a few in the Hartford office, suggesting that the pressure for forming new connections is less in the Hartford office.

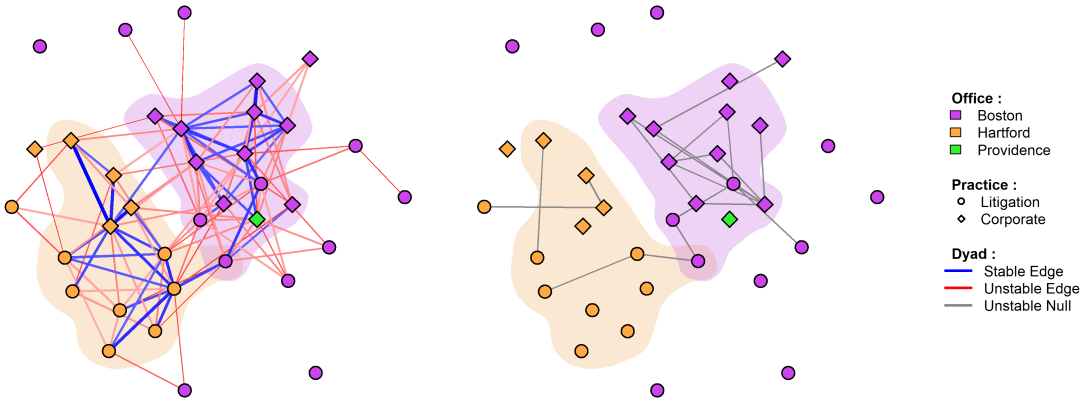

Figure 3.15: Illustration of the Lageza lawyer network with stable/unstable edges and nulls under the model. Specifically, edge stability is illustrated on the left panel, where the usage of color indicates stability. Unstable nulls are plotted on the right panel, with other nulls are stable.

This analysis provides a basis for predicting what changes in the network can be expected. The external unstable edges between two groups suggest that cross-office ties tend to be somewhat fragile, and prone to disruption. The internal unstable edges connecting peripheral partners to core partners within groups likewise suggests an enhanced propensity towards turnover for collaborations involving these more marginal partners. By contrast, the unstable *nulls* among core actors suggest the potential for new collaborations among the most central

partners, especially in the Boston office.

The local stability calculations identify the dyads that are most vulnerable to changes in the network, i.e., if changes *were* to happen, the unstable edges are the ones that would be expected to change the earliest. To see how this corresponds with an explicit dynamic process, we run 100,000 Metropolis trajectories until the first dyad toggle is accepted and record the number of times each dyad was toggled. We plot the fraction of each dyad toggle occurrence as a function of d_i for all dyad D_i. When $d_i < 0$ (left panel of fig.3.16), indicating dyad D_i is stable and the probability of accepting an edge toggle is proportional to the ratio of target/alternative graph potentials. When $d_i > 0$ (right panel of fig.3.16), dyad D_i is unstable and the fraction of dyad toggles becomes flat and is equal to the probability of any dyad being sampled ($\binom{v}{2}^{-1}$, which is 1/630 in this case). This reflects the fact that the Metropolis acceptance probability becomes 1 when a move is favored. This experiment shows that the most vulnerable dyads are the ones that are outside of the stable region, where the probability of such a dyad toggle is equal to the probability of any dyad being sampled. When a dyad lies within the stable region, the probability of a toggle is an exponentially increasing function of negated distance to the closest hyperplane, maxed out at dyad sampling probability. This indicates the dyads lie close to the hyperplanes, although within the stable region, are also somewhat vulnerable to structural changes. It is also interesting to note that under this model there are more unstable edges than nulls, indicate a tendency towards lowering network density by breaking established edges.

3.5 Discussion and Conclusions

In this paper, we proposed a novel approach to analyze graph stability, based on the relative favorability of a target graph vis-à-vis a set of alternative graphs under a probability model. The requirements for applicability of this approach are very general, being merely a target

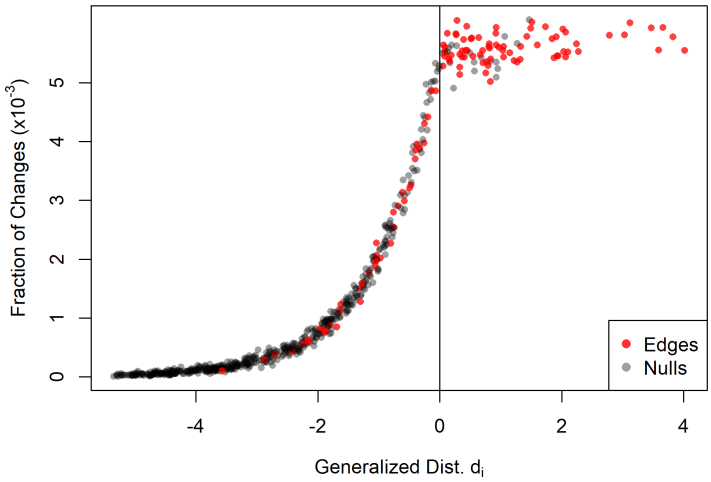

Figure 3.16: Dyad toggle occurrence fraction as a function of generalized distance d_i. Points on the left panel ($d_i < 0$) indicates dyads within the stable region, and points on the right ($d_i > 0$) indicate unstable dyads. Edges and nulls are differentiated by color.

graph, a set of alternative graphs against which stability is to be assessed, and a model class that parameterizes network probability as a monotonically increasing function of the linear combination of parameter values and graph statistics ($\boldsymbol{\theta}^T \mathbf{t}(g)$). ERGMs are a widely used model class of such kind, with the probability of a graph being proportional to the exponentiated linear terms. The stable region, if it exists, is the region where the target graph is more probable than any of the proposed alternative graphs. We construct an $|S|$-by-K matrix M, each row of which is the generalized change score between the target graph and one of the alternative graphs (defined as $t(G') - t(G)$). Under this setting the stable set of a model is $\theta : \mathbf{M}\theta \in \mathbb{R}_{-}^{|S|}$. We show that the stable region (if exists) is the interior of a convex K-polytope cone that points at the origin, and each facet of the cone is a row in M. We show an intuitive and easily implemented method to convert the facet representation to the ridge representation, whose worst-case complexity is the same as the exhaustively searching through all possible intersections. Nevertheless, the best case or the average case complexity is a massive speedup over the exhaustive method.

Although this definition applies broadly to any target graph and arbitrary alternative set,

certain alternative sets are particularly useful for providing intuition regarding model behavior and potential dynamics. For example, if we construct the alternative set to be all graphs that are Hamming distance one from the target graph, then stability against this set indicates that any one-step change is disfavored. For a random walk algorithm that assumes changes happen in series (i.e., one change at a time), this very local form of stability is an easily calculated approximation for dynamic stability (since any longer trajectory must still begin with a single step).

To demonstrate our approach we showed a simple social example - an isolated star graph - inspired by the structures arising in certain charismatic cult groups. Local stability ($S = H^{(1)}$) is particularly immediate in this case, as S includes only two isomorphism classes. Under a minimal model family, we verify that the stable region corresponds closely to the region that remains dynamically stable under random walk MCMC trajectories, with dynamic stability fading as one approaches the facets of the stable region. We further investigate the potential differences of graphs in S near the facet of the stable region and show how dynamic stability is related to the potential difference of the alternative graphs versus the target. When S is chosen to be $H^{(1)}$, any instability w.r.t. S can also be interpreted in terms of dyad vulnerability, that is, under the current model forces, if changes were to happen to the network, which dyad is most vulnerable to being toggled. We divide the parameter space into regions that represent the types of dyads that are most likely to undergo changes from both theoretical derivation and simulation.

We demonstrate a straightforward approach to exploring network stability under a given model, using the Lazega Lawyer dataset as an example. By simply constructing the $\mathbf{M}$ matrix and examining whether the vector $\mathbf{M}\hat{\boldsymbol{\theta}}$ is in the negative quadrant, i.e., $\mathbf{M}\hat{\boldsymbol{\theta}} \in \mathbb{R}_{-}^{|S|}$, we can assess whether the observed graph is predicted to be stable under the estimated model. When the graph is not locally stable, then we may further inquire into the stability of particular dyads (exploiting the relationship between dyad toggles and the elements of S). Thus, non-

negative elements in $\mathbf{M}\hat{\boldsymbol{\theta}}$ suggest the instability of the corresponding dyads. This can aid in making predictions regarding potential future changes to the network under the current social forces (as parameterized by the estimated model). For the Lazega dataset, we are able to identify the structural characteristics of the network from dyad stability assessment. For example, by examining the stable edges we are able to identify two densely connected clusters centered at two of the larger offices; each exhibits a core-periphery structure. The unstable nulls are relatively concentrated in the Boston office, suggesting that the pressure to collaborate is greater in the Boston office, while unstable edges bridging offices suggest fragility in ties between units. Such insights may be useful for guiding additional empirical studies or modeling efforts.

The methodology introduced here offers a powerful new toolset for practitioners of network modeling. The techniques presented require no simulation, and are applicable to a wide range of problems. The one-step stability metric introduced herein is a straightforward implementation of the alternative set, and possess an intuitive interpretation. At the same time, our method is amenable to any user-defined alternative set. Our approach also offers a quantitative tool for measuring instability in network structure due to either a strain imposed on a network by the social forces at play, or conversely, could indicate that the choice of social forces in the model could be poorly chosen. Finally, we also note that the formal correspondence between the ERGM form and Boltzmann distribution makes this approach useful in physical settings [Grazioli et al., 2019b], where locally stable structures correspond to local energy minima in graph space. The ability to easily characterize the conditions under which particular graph structures are energetically favorable may be useful for studying the formation of complex materials, or the protein aggregates associated with Alzheimer's and other diseases.

Chapter 4

Scalable Estimation for DNR Models of Sexual Contact Networks from Retrospective Life History Data

4.1 Introduction

The analysis of network dynamics – changes in the edge structure of a network – has been of theoretical interest to sociology for a number of years. With the advances in modern computational technologies, recent work on analyzing dynamic social networks has been putting a strong emphasis on formal statistical modeling [Watts, 2004]. More powerful computers and cheaper storage servers have also allowed many large-scale electronic datasets to be available, which intensifies the need for salable modeling tools. In this work, we focus on analyzing sexual contact networks (SCNs), which is one of the central topics in disease spreading studies, that requires large-scale estimation to model the complex interactions between actors.

Sexually transmitted infections (STI) represent a significant public health problem both in the United States and worldwide. A 2013 study [Satterwhite et al., 2013] estimated that in 2008, there are 110 million prevalent sexually transmitted infections in the United States, and 20 million new infections occur every year. WHO's report [Organization et al., 2015] estimated that in the same year, the new infections of four curable STIs - Chlamydia trachomatis, Neisseria gonorrhoeae, syphilis, and Trichomonas vaginalis - reached 499 million which is 11.3% higher than 2005.

The measurement and modeling of large-scale sexual contact networks are central to the prediction and prevention of STI spread [Foulkes, 1998, Anderson and Garnett, 2000]. Indeed, the macro structure of SCNs has been shown to predict how quickly HIV/STIs can spread through a network [Morris, 1993b]. Additionally, the local SCN structure around an actor helps explain his/her risk of acquiring or transmitting HIV/STIs [Kretzschmar, 2000, Ghani and Garnett, 2000].

Currently, models predicting the structure of sexual contact networks are limited by computation. Though the exponential family of models can dynamically model sexual contact networks at small scales, temporal exponential random graph models (TERGM) applied to large SCNs from sampled data require simulation of the entire population, which is computationally intractable on the order of 10^4 nodes. If researchers want to model large-scale sexual contact networks, more efficient methods will need to be introduced.

In this chapter, we present a method that draws inference from sampled data utilizing dynamic network logistic regression [Almquist and Butts, 2014]. With this method, we no longer need to simulate or model the full network to obtain the inference of the network coefficients. We quantitatively show that the inference was able to capture the underlying properties of the network. We also show that this method is able to scale to large populations with sample data being almost fixed in size.

4.2 Background

Logistic regression applied to adjacency matrices has a long history in the discipline [eg. Krackhardt, 1987, 1988], and is a natural analysis choice if edge-wise independence is assumed. Almquist and Butts [2014] have built upon this work and well-used approach of exponential random graph models, commonly known as ERGM [Butts, 2008a, Snijders, 2002, Strauss and Ikeda, 1990], to formally introduce dynamic network logistic regression (DNR), the implementation of which has drastically reduced complexity compared with the traditional ERGMs and their subfamily, separable temporal ERGMs (STERGMs, Krivitsky and Handcock [2014]). DNR is equivalent to logistic regression with lagged covariates, so it uses a familiar method that is easily implemented using a number of software applications. Generally, when predicting elements of large social networks, DNR is preferable computationally, as DNR is much more scalable [Komarek and Moore, 2003].

Given the scalability problem of STERGM [Leifeld and Cranmer, 2019], DNR has obvious applications to large scale network analysis. A widely studied example of such networks is sexual contact networks (SCNs), which is a group of persons who are connected to one another sexually. A common practice in SCNs is to use vertices to represent individuals and edges to represent sexual relationships. SCN studies often depend upon large-scale population structures and dynamics. For example, Morris et al. [2009] use concurrency behavior (described as having two or more sexual partnership at a given time), combined with population dynamics to study the race disparity in HIV prevalence. Modeling large-scale network dynamics are thus useful in predicting the spread of HIV and other sexually transmitted infections in large populations, so the use of scalable methodology is of great use to this subfield.

Compared with many other social networks, SCNs possesses some unique features: Most nodes in an SCN are tied to one or less nodes at any time (i.e., most people do not maintain

more than one sexual partner simultaneously); and it generally has a bipartite-like structure due to the strong heterophily on sexual ties (i.e., most relationships are heterosexual); finally, sexual contacts generally churn slowly relative to the scale of measurement (which is stated in the next paragraph): National Health and Social Life Survey (NHSLS) collected data suggests that the average tie length being on the order of ten or so years in the United States [Laumann, 1994].

Measuring sexual contacts among a population has commonly relied on sampled data using either prospective or retrospective life history designs. Among the two, retrospective life history (RLH) designs are better suited for large-scale measurement over a long period of time [Reading, 1983, Leigh et al., 1998, Tran et al., 2013]. For practical reasons, a population-sized RLH measurement on say, the United States, is not feasible at scale. All currently available datasets of such networks focus on a relatively small samples. E.g., NHSLS [Laumann, 1994], being one of the rather extensive RLH surveys, measured about 3,500 individuals only. In addition to a population sample, life history designs generally measure sexual contacts on a fine-scale (usually monthly), which is as accurate as we can expect humans to be when asked about often imprecise start points for relationships that perhaps began years before an interview.

The edge-wise independence assumption is essential to network analysis with DNR models: For a dynamic network that evolves slowly relative to the time scale of the measurement, dependence among edges is largely captured by a combination of covariate effects and the past history of the network. Imagine a process that samples the network longitudinally; if it samples the network finely enough, such that no more than one edge change could occur in each slice, then edges within the same slice are independent of each other given the covariates and the past slices. SCNs are commonly sampled in a monthly resolution, which is precise relative to the evolution of the network; it is reasonable to assume the edge dependency is entirely captured by the covariates and the past states of the network. These

assumptions make it ideal for modeling SCNs with the DNR framework, a special case of the temporal exponential random graph (TERGM) that is computationally equivalent to logistic regression with lagged independent variables.

During our discussion, we employ the following terminology in describing the measurement, data, and analysis. A participant or subject of an SCN RLH measurement is referred to as an *ego*, and each of his/her sexual partners is called an *alter*; every other individual in the population, that is not an ego nor an alter, is called a *non-alter*. Their attributes are referred to as *ego covariates*, *alter covariates* and *non-alter covariates*. A *dynamic network* is a collection of vertices and their activities (e.g., establishing a connection or breaking an existing connection) over a period of time, and *cross-sectional network* (or *network* for short unless other-wise specified) is a snapshot of the dynamic network to capture the state of the network at a specific time. A dyad is an induced subgraph of two nodes of any relations; if they are connected, then we say there is an *edge* between them, or otherwise, a *null*. With these terminologies and the above independence assumptions in mind, we outline the statistical theory involved with scalable modeling of sexual contact networks using DNR.

4.3 Dynamic Network Logistic Regression

4.3.1 Likelihood Calculation

With ERGM's definition in 1.1, the probability of y can be written as

$$Pr(Y = y|\theta) = \frac{\exp(\theta^\top S(y))}{\sum_{y'} \exp(\theta^\top S(y'))},$$

and the change score Δ_{ij} can be written as

$$\Delta_{ij} = S(y_{ij}^{+}) - S(y_{ij}^{-}).$$

Note that the odds of edge ij being present can also be written in terms of the change score,

$$\begin{aligned}\frac{Pr(Y_{ij} = y_{ij}^{+}|\theta)}{Pr(Y_{ij} = y_{ij}^{-}|\theta)} &= \exp\left(\theta^{\top}(S(y_{ij}^{+}) - S(y_{ij}^{-}))\right) \\ &= \exp(\theta^{\top}\Delta_{ij})\end{aligned}$$

Likelihood of θ of a dynamic network is the joint probability of the network instances at all times given the covariates. With the edge-wise independence assumption, if we sample the network finely enough we remove dependencies between all but consecutive sampling slices. Then the likelihood can be written as:

$$\begin{aligned}\mathcal{L}(\theta) &= \prod_{t=1}^{T} Pr(Y_t = y_t|\theta, Y_{<t} = y_{<t}, X) \\ &= \prod_{i,j,t=1}^{N,M,T} \frac{\exp(\theta^{\top}S(y_{ijt}, Y_{<t} = y_{<t}, X))}{\sum_{y_{ijt}} \exp(\theta^{\top}S(y_{ijt}, Y_{<t} = y_{<t}, X))}\end{aligned}$$

Where N is the population size, M is sample size and T is the total observed time steps. To simplify, let $S_{C_t}(\cdot)$ represent the sufficient coefficient under certain conditions: $S_{C_t}(y_{ijt}) = S(y_{ijt}, Y_{<t} = y_{<t}, X)$. For each i, j, t pair, there are only two possible $y_{i,j,t}$ values, $y_{i,j,t} = 1$ or $y_{i,j,t} = 0$. Thus the above equation becomes:

$$\mathcal{L}(\theta) = \prod_{i,j,t=1}^{N,M,T} \frac{\exp(\theta^{\top}S_{C_t}(y_{ijt}))}{\exp(\theta^{\top}S_{C_t}(y_{ijt}^{+})) + \exp(\theta^{\top}S_{C_t}(y_{ijt}^{-}))} \tag{4.1}$$

We can simplify the equation by dividing both the numerator and the denominator by

$\exp(\theta^\top S(y_{ijt}^-, C_t))$ (the detailed steps are shown in the appendix):

$$\mathcal{L}(\theta) = \prod_{i,j,t=1}^{N,M,T} \frac{e^{y_{ijt}\theta^\top \Delta_{ijt}}}{1 + e^{\theta^\top \Delta_{ijt}}}$$

4.3.2 Binning Cases

Change scores Δ_{ijt} can only take a finite set of values. We can group the identical values into classes, and each of them is called a bin. Let B be the number of values Δ_{ijt} can take, and Δ_b be its value in bin b, eg. for all i, j, t, $\exists b$ such that $\Delta_b = \Delta_{ijt}$. We could easily estimate the number of edges and nulls in each bin, and thus further simplify the calculation. Let n_b be the number of cases in bin b; $n_b^{(E)}$ be the number of dyads that have edges between them (eg. $y_{ijt} = 1$) in bin b; and $n_b^{(N)}$ be the number of dyads that do not have edges between them (eg. $y_{ijt} = 0$) in bin b.

$$\begin{aligned}
\mathcal{L}(\theta) &= \prod_{i,j,t=1}^{N,M,T} \frac{e^{y_{ijt}\theta^\top \Delta_{ijt}}}{1 + e^{\theta^\top \Delta_{ijt}}} \\
&= \prod_{b=1}^{B} \left(\frac{e^{\theta^\top \Delta_b}}{1 + e^{\theta^\top \Delta_b}} \right)^{n_b^{(E)}} \left(\frac{1}{1 + e^{\theta^\top \Delta_b}} \right)^{n_b^{(N)}} \\
&= \prod_{b=1}^{B} \frac{\left(e^{\theta^\top \Delta_b} \right)^{n_b^{(E)}}}{\left(1 + e^{\theta^\top \Delta_b} \right)^{n_b^{(E)} + n_b^{(B)}}} \\
&= \prod_{b=1}^{B} \frac{e^{n_b^{(E)}\theta^\top \Delta_b}}{\left(1 + e^{\theta^\top \Delta_b} \right)^{n_b}}
\end{aligned}$$

The log likelihood is,

$$l(\theta) = \sum_{b=1}^{B} \left(n_b^{(E)} \theta^\top \Delta_b - n_b \log(1 + e^{\theta^\top \Delta_b}) \right) \tag{4.2}$$

We simplified the calculation from looping through i, j, k to only looping through b, and it is reasonable to assume $B \ll NMT$ for a large dataset. All we need to know are the number of dyads and the number of nulls in each bin.

4.3.3 Complexity

The computational complexity is substantially reduced using DNR versus other methods achieving a similar end. The traditional TERGM method has complexity $O(N^2T)$, where N is the population size, and T is the number of time steps a longitudinal network is measured at. Our method essentially breaks down into two procedures, binning the case and calculating the likelihood. In the first procedure, we aggregate the population into bins and make a pass of all {ego, bin, time} triplets to calculate the binned change scores counts. The complexity of this procedure is $O(MBT)$, where M is the sampled ego size and B is the number of bins. The likelihood calculation procedure is only $O(B)$ and the total complexity remains $O(MBT)$. The number of bins B is the number of the unique combinations of dyad attributes and it is reasonable to assume that it does not scale with the population size N (for a large enough population, obviously). This method scales only with the sample size rather than the population size, thus has the potential to scale to a very large population.

4.4 Validation Methods

To test whether the method successfully captures properties of large-scale SCNs, we apply a series of validation methods with both synthetic data and a real dataset. Figure 4.1 shows the validation steps. To outline, we first used the network data from Krivitzky's STERG model [Krivitsky, 2012] to draw the DNR parameters, with which we simulate a set of large-scale dynamic SCNs. We show that the simulated network data captures the properties

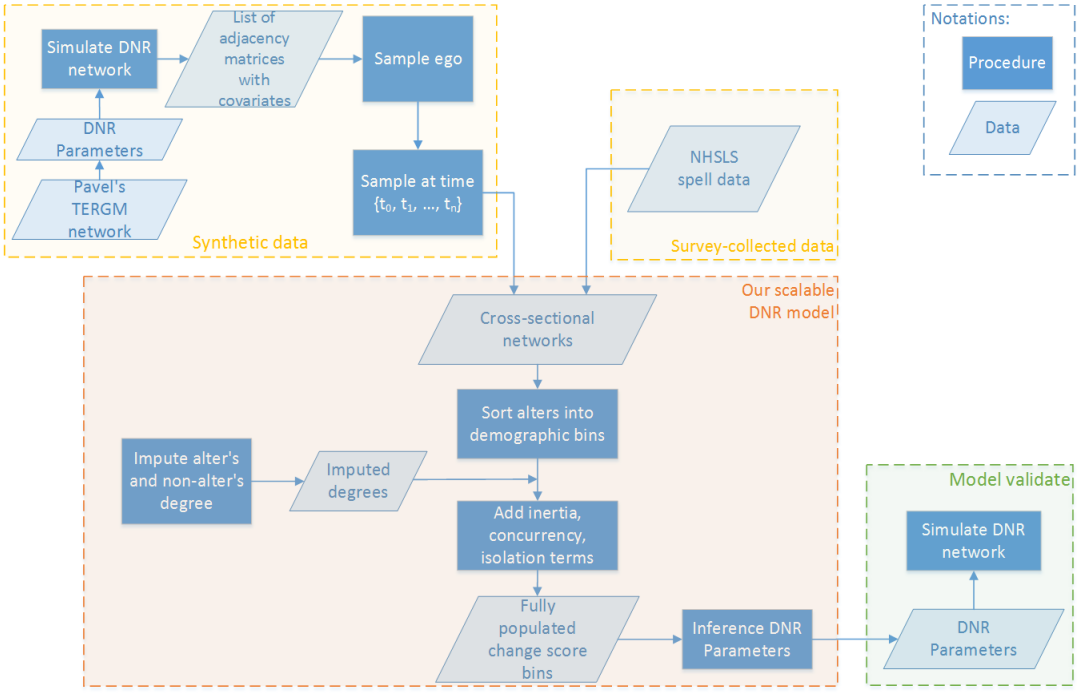

Figure 4.1: System flow and structure for the validation method.

of real-world SCNs. We extract the networks formed by a small sampled population such that they resemble the structures of egocentric networks (sec. 4.4.1). We then re-draw DNR parameters from them and compare them with the DNR parameters drawn from the STERG model (sec. 4.4.2, 4.4.3) . Finally, this procedure is repeated for a real-world dataset of NHSLS [Laumann, 1994].

4.4.1 Synthetic Data Generation

We apply DNR ten times to the network data simulated from Krivitzky's separable temporal exponential random graph model (STERGM) [Krivitsky, 2012] fit to the real-world SCN from NHSLS. Then with the resulting DNR parameters, we simulated synthetic datasets using the DNR model. We compare descriptive statistics from data simulated from the DNR model with that of the STERGM model. Ideally, we should see that the descriptive statistics

simulated with DNR parameters match that of simulated parameters of STERGM.

Following Krivitzky's setup in [Krivitsky, 2012], we simulate a set of 100 large, time-resolved SCNs with population size (N) of 10000; total time points (with monthly resolution; T) of 240 months, and burn-in time (T_b) of 150 months.

We picked six parameters that are commonly used to model SCNs (table 4.1). The mean estimated DNR parameter value is also shown. In the following discussion, we refer to these parameters as the *true parameters*.

Term	Estimate	Description
Intercept	-23.9	Control network density
Gender homophily	-2.3	Model same-gender ties
Race homophily	-1.3	Model same-race ties
Inertia	27.2	Model tie persistence
Concurrence penalty	-0.9	Penalize concurrency
Degree0 penalty	10.5	Penalize isolate nodes

Table 4.1: Parameters and values obtained from TERGM simulated SCNs

Then we sample the network to resemble the egocentric measurement process. We pick $M = 1000$ sampled egos out of population of $N = 10000$, and use all $T = 240$ temporal cross-sections. To assess the effects of M on a fixed population, we also tested the model against $M = 500$. Information on egos' ties, ego covariates, and alter covariates are obtained; alters' ties are recorded but are not always used depends on the experiment setup described in sec. 4.4.2. We refer to this network as *sampled network*, which then becomes the input to the validation model described in sec. 4.4.3.

4.4.2 Concurrency Imputation

Concurrency is a key component to the modeling of SCNs, because it is not only one of the significant indicators of STI prevalence, but also a strong factor influencing tie formation and persistence. However, in most of the egocentric SCN datasets, alter concurrency status

is unknown. This is a substantial missing data problem. If the ego has an alter that is maintaining an additional tie with non-alter, concurrency is present for that alter but not observed in our data set. In NHSLS [Laumann, 1994] there are specific questions that asked egos whether their alter had other sexual relationships besides theirs. Obviously, this measure is limited by the egos' awareness of their partners' sexual ties. Nonetheless, we use this as an indicator that the alter concurrency can practically be assessed by RLH design. To address the missingness caused by the lack of awareness of partners' other sexual partners, and to make sure our model remains robust on SCN datasets that do not have alter concurrency information available, we experiment with many different imputation techniques to recover the missing concurrency values. Specifically, we aim to predict which alters are likely to have an additional tie, and also the concurrency status of non-alters, which is also of consequence to tie formation.

Based on table 4.1 and equation 4.2 the essential information to the DNR model are ego, alter and non-alter covariates; ego-alter ties; and alter & non-alter concurrency. For the purpose of our experiment, it is reasonable to assume ego & alter covariates and ego-alter ties are always available, and detailed non-alter covariates and non-alter concurrency are always unavailable (unless the dataset is large enough to cover the entire population of interest - which is not feasible for large scale SCNs). To remedy the missing non-alter information, since the data we used is based upon a representative sample of the United States, we can borrow from other sources of US-representative data (such as the U.S. Census) in our procedure. For the rest of the information (alter and non-alter concurrency), we test 4 different setups with different levels of missingness. Known everything (KE): This is a control experiment where we assume all information is available. Impute non-alter concurrency (INA): Alter concurrency is perfectly available and we impute non-alter's concurrency. Impute everything (IE): Both alter and non-alter concurrency are unknown and imputed. Impute everything with survey assistance (IEwS): We use alter concurrency results from NHSLS with the same missingness level. The missing alter concurrency and all non-alter concurrency are imputed.

We use standard machine learning tools to impute the alter concurrency. Two classifiers are tested based on the prediction accuracy: support vector machine (SVM) and random forest. SVM being a widely used approach in predicting binary cases, is well suited for modeling complicated interactions between covariates. Random forest is less vulnerable to model turning and over-fitting and is generally less computational costly. The ego and alter covariates together with ego's partnership information (tie duration, concurrency status) used as features to train the classifiers. SVM over-perform random forest with an accuracy of 0.88 (the accuracy of the random forest classifier is 0.77), thus is chosen to predict alter concurrency.

Non-alter concurrency presented in equation 4.2 (eg. $n_b \log(1 + e^{\theta^\top \Delta_b})$) can be calculated by counting the number of dyads in the change score bins. Unlike alter concurrency, we do not need to impute non-alters' individual concurrency status. We simply take the concurrency statistics drawn from NHSLS match with census demographics as the aggregated non-alter concurrency.

4.4.3 Parameter Recovery

The input to the DNR model (described in section 4.3.1) are the sampled network cross-sections (described in section 4.4.1), together with the concurrency data (described in section 4.4.2). Note that a part of the concurrency can be unobserved, and we would use the imputation method (described in section 4.4.2) to predict the concurrency level. For this experiment, we tested four concurrency missingness levels (KE, INA, IE, and IE). We then use the binning method described in 4.3.2 to create a data structure that includes edge toggle from all network cross-sections. We then classify the toggles into bins, such that each bin of properties has a number associated with them, which is the number of toggles belong to this bin. With this data structure, we estimate the DNR parameters from each input 100

times and compare them with the true DNR parameters that we simulate the full networks from. Ideally, the true parameters should lie within the scope of the 95% confidence interval of the estimates.

Additionally, to show that we are able to apply this method to a much larger dataset, we use a real-world longitudinal sexual contact data obtained by NHSLS as the sample data, and the U.S. census as the entire population. We take the 3432 respondent as sampled egos and use the 12-month detailed sexual activity record to form a time aggregated spell set. Similar to the procedure described in 4.4.2, we impute missing alter concurrency and all non-alter concurrency. Then we process the data by binning the change score into unique classes, then apply DNR with the same parameterization (as shown in table 4.1) to the binned data set.

4.5 Results

4.5.1 Application to STERGM-Simulated Data

The results of the DNR parameters estimated from the sampled network are shown in fig. 4.2 (and the exact numbers are shown in the table). We were able to successfully recover most of the STERGM parameters used in the simulation with some certainty, given all but two having 95% CI cover the true value. More specifically, in (a), which we assume both alter and non-alter currency information is known, we are able to get very accurate estimate of the true parameter. In (b) we only impute non-alter degrees, which is the same set up as in NHSLS, we are also able to recover the parameters with very low bias. (c) shows the case we impute both alter degree and non-alter degree. Except for intercept and race homophily terms, the rest have the true values within the 95% CIs. Obviously, in (c) we use the least amount of information, and the estimates indeed deviate from the true value with the largest bias among 4 cases. Nonetheless, the biases are still very insignificant compared to the true

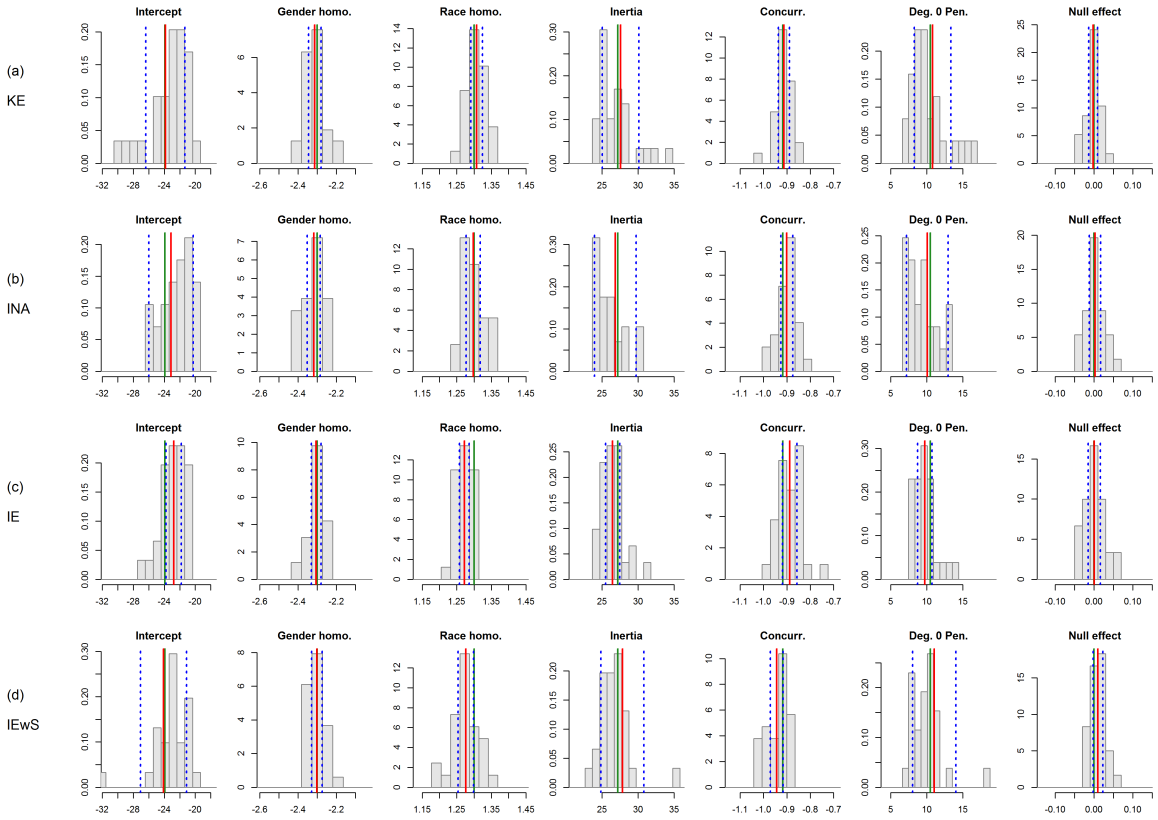

Figure 4.2: The result of the DNR parameter estimation from the sythetic network samples. 95% CI covers all but two true parameter values.

parameter values. In (d) we use NHSLS's alter degree missing rate as a basis and only impute the missing alters. With this remedy, we are able to recover the biased estimated from case (c).

To investigate the effects of the sample size on fixed population, we experiment with a smaller sample size $M = 500$, with the rest of the settings holding unchanged. 2 out of the 4 missing levels are tested: (a) KE and (d) IEwS, among which the former is an ideal setup with no missingness, and the latter is a practical setup with the remedy of surveyed alter concurrency.

Term & True Val.		Bias				Standard Error			
		KE	INA	IE	IEwS	KE	INA	IE	IEwS
Intercept	-23.9	0.043	0.770	**1.123**	-0.189	1.273	1.448	**0.498**	1.505
Gender homo.	-2.3	-0.012	-0.017	-0.005	-0.001	0.016	0.017	0.013	0.014
Race homo.	1.3	0.007	-0.002	**-0.028**	-0.024	0.009	0.010	**0.007**	0.011
Inertia	27.2	0.367	-0.362	-0.736	0.632	1.277	1.459	0.501	1.508
Concurr.	-0.92	0.007	0.020	0.032	-0.024	0.012	0.013	0.016	0.014
Deg. 0 Pen.	10.5	0.303	-0.427	-0.765	0.541	1.277	1.456	0.502	1.510
Null effect	0	-0.002	0.003	0.001	0.011	0.006	0.007	0.008	0.006

Table 4.2: DNR parameters estimated from sampled network. The true value of the parameters are shown in the parenthesis. The terms are intercept, gender homophily, race homophily,inertial, concurrence penalty, degree 0 penalty and null effect.

4.5.2 Application to NHSLS

We fit the DNR model to an egocentric SCN dataset extracted from the NHSLS. The survey has detailed demographic information as well as respondents' sexual activities for the past year (prior to the interview date). We take all respondents in the survey as the sampled egos (total 3432 respondents), and form a time aggregated set of 12 months. The alter concurrency status is taken directly from the survey, unless the respondent left it empty. The missing alter concurrency information (with 15.91% missing rate) is imputed with the same methods used in the synthetic data. Although from a computational standpoint, the population can be set to an arbitrary size, it is reasonable to set it to the coverage of the survey. NHSLS was conducted in 1992 within the United States and sampled respondents age 18 to 59. We restrict our population to be within the same scope, with the census estimates being 145.5 million [Byerly and Deardorff, 1995]. We then fit 1000 NDRs to this time aggregated spell dataset and with the same model terms as shown in 4.1. The resulting coefficients are shown in table 4.3.

	Intercept	Gender homo.	Race homo.	Inertia	Concurr.	Deg. 0 pen.
Mean	-35.95	-19.58	20.89	50.79	-19.54	-230.77
S.E.	2.96	0.18	0.33	1.18	0.37	5.18

Table 4.3: Estimated DNR coefficients of the NHSLS spell data, with the population covers the entire the United States.

4.6 Discussion and Conclusions

Above, we show that we could successfully draw inferences from a large sexual contact network using a sampled population of egos. Though our sample size ($M = 1000$) equals only 1/10 of the entire population ($N = 10000$), our results show relative small bias, with good statistical power (low standard errors; the true value lies within 95% CIs of the estimates). Sexual contact networks are sufficiently sparse to allow for assumptions regarding independence (i.e., independence of ties conditioned on covariates and the past network status), and the time scale in which SCNs are measured allow for researchers to observe network dynamics in precise time scales. We have shown that these assumptions indeed allow for computationally efficient and accurate estimation of parameters from a large network.

Central to the estimation is the use of imputation to recover missing values implicit in data collection. Using Census demographics and sample characteristics, alter's and non-alter's concurrency was imputed relatively accurately for DNR estimation, with a few exceptions. Though common data collection techniques do not provide all the information necessary for a dynamic network logistic regression model to be fit, our imputation procedures provide a remedy with which we are able to obtain estimates with low bias and high precision. Nonetheless, we show that with the aid of specific survey questions (i.e., awareness of alter concurrency), though imperfect, provides a good basis for the imputation method to improve.

State of the art dynamic network analysis has shown to be computationally difficult with large networks. For cases where the estimation of a large network is critical in answering theoretical questions or making predictions about diffusion dynamics, this is a significant

cost. Here, we show that we can efficiently estimate the properties of large networks with some dependence assumptions. Comparing with the traditional TERGM method with a complexity of $O(N^2T)$, our method only requires $O(MBT)$, which is much smaller for large scale networks. In addition, since it scales with the size of the sample rather than the population, and we show that the sufficient sampled population size does not necessarily grow with the size of the population, our method can potentially scale to arbitrarily large populations.

Though the above DNR method gives decent estimates for most of the situations, many improvements could be applied throughout of process. We used about 50 representative features for the concurrency imputation classifier to achieve a reasonable accuracy, which is far less than the 1600 attributes in the NHSLS. With this extensive amount of information the survey provided, it is a perfect ground to investigate more advanced machine learning tools. In the chapter, we briefly tested the performance of different sample sizes on a fixed population, although more attention should be drawn to it. It would be very interesting to test whether there is a sufficient M (combined with T) for different population sizes, providing a practical basis for researchers to design future measurement and analysis methods.

Sexual contact networks are an important information source for the diffusion of STI, which is a significant public health problem in the US. Our application revolves around sexual contact networks as a case where the dependence assumptions can ostensibly meet in most cases, and diffusion dynamics are often predicted using large-scale network dynamics. Thus, our method provides a way to efficiently estimate SCN properties with low computational cost, broadening our ability to make predictions are large scales.

Chapter 5

Conclusion and Future work

In this thesis, I have addressed several key questions raised in the field of complex time-varying networks: Measurement, modeling, and computation. The first question we address lies within the realm data collection process. Data collection is often the first challenge researchers face. With the advance in modern technologies and the development in computational resources, storage space became the least constraint. Nevertheless, other aspects still rule the amount, the quality and the cost of the data collection. For example, some data collection process requires human involvement which could largely bring up the cost. Meanwhile, OSN data can be collected automatically; however, it is subject to each OSN platforms' own regulation. Essentially, how much data, and what data one could access almost entirely depends on OSN platforms' API designs. With these limitations imposed, some portion of the data is unavoidably missing, how does it impact the network data we collected? If possible, are there designs that are intrinsically better than other designs? The second question we address is about the fundamentals of dynamics; that is, what mechanisms drive the changes in a network? The third question we address is about the scalability of models for dynamic networks. While smaller networks reveal delicate relations of actors, large scale inference is often necessary or desired. Traditional methods suffer when networks

grow in size, and we aim to answer whether there are methods that approximate the inference well, and at the same time, tractable when networks are large.

With the first question in mind, in chapter 2, we characterized two designs that are widely used in RLH collection–intervalN and lastK. We show the key differences between the two designs and their variants. For this study, we apply the designs on an extensive dataset, NHSLS, and examine their impact. We show that intervalN design, although having niche advantages in certain situations, rarely over-performs the lastK design, thus is not recommended when the purpose of the data collection is rather general. We also quantitatively prove that between the two subdesigns of the lastK regime, the terminal selection is almost always superior to its alternative, the onset selection. We quantitatively examine the missingness caused by each design, and how it further impacts the modeling and imputation of such networks. In this chapter, we also provide some insights for researchers when deciding the design schemes that are most suitable for their studies.

Chapter 3 makes contributions to the second question to which we approach from a novel direction. Instead of asking the question "what mechanisms drive the changes of graphs", we ask, "what mechanisms anchor graphs in place?" We assess this by proposing a set of alternative graphs, which we anchor our target graph against. We prove that for any models that characterize the probability of networks as a monotonically increasing function of the linear combination of the parameter values and the graph statistics, then the stable region in the parameter space, if exists, a convex polytope pointed at the origin. We propose data structures and algorithms that simplify the computation of the stable region, and show its computational advantages. We apply this method to the Lazega lawyers dataset and analyze the parametric distance to the stable region when toggling each edge. We show that the distance to the stable region is a functional form of how likely the edge being toggled. This finding not only assists us on predicting which edges are vulnerable to be altered, but also sheds light on the question "what forces drive the graph to evolve over time".

Chapter 4 answers the last question with a scalable approximation to the network modeling parameters. For a network of size n and with t time-dependent snapshots, the complexity of the traditional approach is $O(n^2t)$. We proposed an algorithm for the approximate approach - dynamic network regression (DNR), which has a much manageable complexity $O(nmt)$, where m is the sample size (much smaller than n). We show that with the reduction in computational complexity, the parameter approximation is statistically acceptable under different levels of missingness.

This thesis not only makes attempts to answers the three major questions, but it also provides many key findings related to time-varying systems. Among them, I include four broad takeaways that improve our understanding of such system.

- In an ideal world where there were no limitations on means of data collection, privacy, storage space, and handling accuracy, researchers would gather every single piece of information that is useful. In practice, there will be *unavoidable* missingness in the dataset. It is the researchers' responsibility to understand the limitation, thus correctly estimate how much power they have when making assertions from such data. It is also important for researchers to choose the suitable data collection design, in order to maximize the useful data to cost ratio.

- For retrospective spell data collection, when it relies on human subjects to recall, report, and describe the past spell events, it is best to collect the ones that ended the most recently. Not only do the subjects have the freshest memory of them, but also such spells tend to give less missingness retrospectively.

- The strength of a connection under certain social forces is well studies; however, the strength of a null (i.e., non-connection) is often neglected. An intuitive approach is that we could toggle the null (i.e., creating an edge between dyads of interests), and examine the strength of the newly created edge, the strength of that edge is a reasonable

(inverse) indicator of how strong the null is.

- For models that focus on countable properties of the social networks (e.g., number of occurrences), we often see a large number of repeated entries. This should provide researchers a good intuition for improving the scalability of the model. One could bin the entries with exactly the same properties, and record the number of times that entry appears in the dataset.

We also face several challenges that limit certain progress along with the research. This list summarizes the challenges, and hopefully provides researchers with a clear direction for improvement.

- The stable region approach, mentioned in chapter 3, is ideal for studies that focus on the absolute structure of the network, e.g., the cult structure. Although we intuitively believe there must be social structures that favor a certain state, and might collapse after a small number of changes, we have not yet found such data sets in the literature.

- Most SCN datasets do not provide information on alter's concurrency, which is crucial to many disease transmission studies. Though our DNR imputation procedures (discussed chapter 4) provide a remedy, we show that with the aid of specific survey questions (i.e. awareness of alter concurrency), though imperfect, provides a good basis for the imputation method to improve.

As a final note, the results presented herein suggest a series of potential future directions. We believe these directions are to the interest of researchers studying time-varying networks, and serve the purpose of advancing the field of as a whole.

- In chapter 2 we discussed one source of missingness caused entirely by the designs. Many other sources of missingness (e.g., subjects' memory loss, or unwillingness to

provide information) and error (e.g., handling error) may also occur. A question for further research is how these other sources of error or missingness interact with the design effects studied here.

- In chapter 2, The ground truth network is a synthetic population case, calibrated to reproduce the main features of the original data set on which it is based. We do not account for demographic effects. Although we believe it is unlikely to alter the results of primary interest for our study, we look forward to follow-up studies using more elaborate models, incorporating more subtle effects, if and when those become available.

- The stability approach to the network dynamics is an elegant way to analyze what forces change or anchor the network. Nevertheless, social networks, being one of the most affluent sources of network data, are often fluid. Researchers are often interested in a direction along which a series of changes happens. One could expand the definition of stable region to stable direction or stable path - that is, changes along the target path are always more likely than off-direction changes.

Bibliography

Daron Acemoglu, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. Systemic risk and stability in financial networks. *American Economic Review*, 105(2):564–608, 2015.

Adaora A. Adimora and Victor J. Schoenbach. Social Context, Sexual Networks, and Racial Disparities in Rates of Sexually Transmitted Infections. *The Journal of Infectious Diseases*, 191(Supplement):S115–S122, 02 2005. ISSN 0022-1899. doi: 10.1086/425280.

Zack W. Almquist. Random errors in egocentric networks. *Social Networks*, 34(4):493 – 505, 2012. ISSN 0378-8733. doi: https://doi.org/10.1016/j.socnet.2012.03.002.

Zack W Almquist and Carter T Butts. Logistic network regression for scalable analysis of networks with joint edge/vertex dynamics. *Sociological methodology*, 44(1):273–321, 2014.

Roy M Anderson and Geoffrey P Garnett. Mathematical models of the transmission and control of sexually transmitted diseases. *Sexually transmitted diseases*, 27(10):636–643, 2000.

David Avis and Komei Fukuda. A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. *Discrete & Computational Geometry*, 8(3):295–313, 1992.

C. Bradford Barber, David P. Dobkin, David P. Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483, 1996.

H. Russell Bernard, Peter Killworth, David Kronenfeld, and Lee Sailer. The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology*, 13(1):495–517, 1984.

Scott A. Boorman. A combinatorial optimization model for transmission of job information through contact networks. *Bell journal of Economics*, 6(1):216–249, 1975.

Tom Broekel and Marcel Bednarz. Disentangling link formation and dissolution in spatial networks: An application of a two-mode stergm to a project-based r&d network in the german biotechnology industry. *Networks and Spatial Economics*, 18(3):677–704, 2018.

Carter T Butts. Social network analysis: A methodological introduction. *Asian Journal of Social Psychology*, 11(1):13–41, 2008a.

Carter T. Butts. network: a package for managing relational data in R. *Journal of Statistical Software*, 24(2), 2008b.

Carter T. Butts. Social network analysis with sna. *Journal of Statistical Software*, 24(6), 2008c.

Carter T. Butts. Bernoulli graph bounds for general random graphs. *Sociological Methodology*, 41:299–345, 2011.

Edwin Byerly and Kevin Deardorff. *National and State Population Estimates, 1990 to 1994*, volume 1127. Bureau of the Census, U.S. Government Printing, 1995. URL `https://www.census.gov/prod/1/pop/p25-1127.pdf`.

Duncan S. Callaway, Mark E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. Network robustness and fragility : Percolation on random graphs. *Physical review letters*, 85(25): 5468, 2000.

Nicole Bohme Carnegie, Pavel N. Krivitsky, David R. Hunter, and Steven M. Goodreau. An approximation method for improving dynamic network model fitting. *Journal of Computational and Graphical Statistics*, 24(2):502–519, 2015. doi: 10.1080/10618600.2014.903087.

Robin Cowan and Nicolas Jonard. Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control*, 28(8):1557–1575, 2004.

Winston Davis. Heaven's Gate: A study of religious obedience. *Nova Religio: The Journal of Alternative and Emergent Religions*, 3(2):241–267, 2000.

Edith D. De Leeuw. Reducing missing data in surveys: An overview of methods. *Quality & Quantity*, 35(2):147–160, 2001.

Peter Sheridan Dodds, Duncan J. Watts, and Charles F. Sabel. Information exchange and the robustness of organizational networks. *Proceedings of the National Academy of Sciences*, 100(21):12516–12521, 2003.

Irene A. Doherty, Nancy S. Padian, Cameron Marlow, and Sevgi O. Aral. Determinants and Consequences of Sexual Networks as They Affect the Spread of Sexually Transmitted Infections. *The Journal of Infectious Diseases*, 191(Supplement):S42–S54, 02 2005. ISSN 0022-1899. doi: 10.1086/425277.

Jane Elliott. *Using Narrative in Social Research: Qualitative and Quantitative Approaches*, chapter 4. Sage, 2005.

Bob Erens, Sally McManus, Julia Field, Christos Korovessis, AM Johnson, Kevin Fenton, and Kaye Wellings. *National Survey of Sexual Attitudes and Lifestyles II: Technical Report*. National Centre for Social Research, London, 2001.

Bob Erens, Andrew Phelps, Soazig Clifton, Catherine H Mercer, Clare Tanton, David Hussey, Pam Sonnenberg, Wendy Macdowall, Nigel Field, Jessica Datta, et al. Methodology of the third British national survey of sexual attitudes and lifestyles (NATSAL-3). *Sexually Transmitted Infections*, 90(2):84–89, 2014.

Bethany G. Everett, Katharine F. McCabe, and Tonda L. Hughes. Sexual orientation disparities in mistimed and unwanted pregnancy among adult women. *Perspectives on Sexual and Reproductive Health*, 49(3):157–165, 2017. doi: 10.1363/psrh.12032.

David L. Featherman. *Retrospective Longitudinal Research: Methodological Considerations*. Center for Demography and Ecology, University of Wisconsin-Madison, 1979.

Kay B. Forest, Phyllis Moen, and Donna Dempster-McClain. The effects of childhood family stress on women's depressive symptoms. *Psychology of Women Quarterly*, 20(1):81–100, 1996.

Mary A Foulkes. Advances in hiv/aids statistical methodology over the past decade. *Statistics in Medicine*, 17(1):1–25, 1998.

Ove Frank and David Strauss. Markov graphs. *Journal of the American Statistical Association*, 81:832–842, 1986.

Deborah Freedman, Arland Thornton, Donald Camburn, Duane Alwin, and Linda Young-DeMarco. The life history calendar: A technique for collecting retrospective data. *Sociological Methodology*, pages 37–68, 1988.

Azra C Ghani and Geoffrey P Garnett. Risks of acquiring and transmitting sexually transmitted diseases in sexual partner networks. *Sexually transmitted diseases*, 27(10):579–587, 2000.

Gianmarc Grazioli, Rachel W. Martin, and Carter T. Butts. Comparative exploratory analysis of intrinsically disordered protein dynamics using machine learning and network analytic methods. *Frontiers in Molecular Biosciences, Biological Modeling and Simulation*, 6 (42), 2019a. doi: 10.3389/fmolb.2019.00042.

Gianmarc Grazioli, Yue Yu, Megha H. Unhelkar, Rachel W. Martin, and Carter T. Butts. Network-based classification and modeling of amyloid fibrils. *Journal of Physical Chemistry, B*, 123(26):5452–5462, 2019b. doi: 10.1021/acs.jpcb.9b03494.

Edward C. Green, Daniel T. Halperin, Vinand Nantulya, and Janice A. Hogle. Uganda's HIV prevention success: the role of sexual behavior change and the national response. *AIDS and Behavior*, 10(4):335–350, 2006. doi: 10.1007/s10461-006-9073-y.

Karen Benjamin Guzzo. New partners, more kids: Multiple-partner fertility in the United States. *The Annals of the American Academy of Political and Social Science*, 654(1): 66–86, 2014. doi: 10.1177/0002716214525571.

Deven T. Hamilton and Martina Morris. The racial disparities in STI in the U.S.: Concurrency, STI prevalence, and heterogeneity in partner selection. *Epidemics*, 11:56 – 61, 2015. ISSN 1755-4365. doi: https://doi.org/10.1016/j.epidem.2015.02.003.

Deven T. Hamilton, Mark S. Handcock, and Martina Morris. Degree distributions in sexual networks: a framework for evaluating evidence. *Sexually transmitted diseases*, 35(1):30, 2008.

Mark S. Handcock and Krista J. Gile. Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5, 2010.

Mark S. Handcock and James H. Jones. Likelihood-based inference for stochastic models of sexual network formation. *Theoretical Population Biology*, 65(4):413–422, 2004.

Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, 24(1):1–11, 2008.

Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky, and Martina Morris. *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (`https://statnet.org`), 2019. URL `https://CRAN.R-project.org/package=ergm`. R package version 3.8.0.

Kathleen M. Harris, Carolyn T. Halpern, Eric Whitsel, Jon Hussey, Joyce Tabor, Pamela Entzel, and J Richard Udry. The national longitudinal study of adolescent health: Research design. *Available at http://www.cpc.unc.edu/projects/addhealth/design*, 2009.

Robert M Hauser, Shu-Ling Tsai, and William H. Sewell. A model of stratification with response error in social and psychological variables. *Sociology of Education*, pages 20–46, 1983.

Paul W. Holland and Samuel Leinhardt. Transitivity in structural models of small groups. *Comparative Group Studies*, 2:107–124, 1971.

Chih-Scheng Hsieh, Jaromir Kovářík, and Trevon Logan. How central are clients in sexual networks created by commercial sex? *Scientific Reports*, 4(7540), 2014. doi: 10.1038/srep07540.

Michel Hubert and Nathalie Bajos. *Sexual Behaviour and HIV/AIDS in Europe. Comparisons of National Surveys*. 1998.

David R. Hunter. Curved exponential family models for social networks. *Social networks*, 29(2):216–230, 2007.

David R Hunter and Mark S Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583, 2006.

David R. Hunter, Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3), 2008.

David R. Hunter, Pavel N. Krivitsky, and Michael Schweinberger. Computational statistical methods for social network analysis. *Journal of Computational and Graphical Statistics*, 21:856–882, 2012.

Jerry A. Jacobs and Rosalind B. King. Age and college completion: A life-history analysis of women aged 15-44. *Sociology of Education*, pages 211–230, 2002.

AM Johnson, J Wadsworth, Kaye Wellings, and Julia Field. *The National Survey of Sexual Attitudes and Lifestyles*. Blackwell Scientific Press, Oxford, 1994.

Doyle P. Johnson. Dilemmas of charismatic leadership: The case of the People's Temple. *Sociological Analysis*, 40(4):315–323, 1979.

Gunnar W. Klau and René Weiskircher. Robustness and resilience. In Ulrik Brandes and Thomas Erlebach, editors, *Network Analysis: Methodological Foundations*, chapter 15, pages 417–437. Springer-Verlag, Berlin, 2005.

Paul Komarek and Andrew W Moore. Fast robust logistic regression for large sparse datasets with binary outputs. In *AISTATS*, 2003.

David Krackhardt. Cognitive social structures. *Social networks*, 9(2):109–134, 1987.

David Krackhardt. Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social networks*, 10(4):359–381, 1988.

David Krackhardt and R. N. Stern. Informal networks and organizational crises: An experimental simulation. *Social Psychology Quarterly*, 51:123–140, 1988.

Mirjam Kretzschmar. Sexual network structure and sexually transmitted disease prevention: a modeling perspective. *Sexually transmitted diseases*, 27(10):627–635, 2000.

Pavel N. Krivitsky. Modeling of dynamic networks based on egocentric data with durational information. *Technical Reports and Preprints; Pennsylvania State University Department of Statistics*, 12(01):1–32, 2012.

Pavel N Krivitsky and Mark S Handcock. A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):29–46, 2014.

Pavel N. Krivitsky, Mark S. Handcock, and Martina Morris. Adjusting for network size and composition effects in exponential-family random graph models. *Statistical Methodology*, 8(4):319–339, 2011.

Edward O Laumann. *The social organization of sexuality: Sexual practices in the United States*. University of Chicago Press, 1994.

Edward O. Laumann, John H. Gagnon, Robert T. Michael, and Stuart Michaels. *National Health and Social Life Survey, 1992*. Inter-university Consortium for Political and Social Research, 1996.

Emmanuel Lazega. *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*. Oxford University Press on Demand, 2001.

Philip Leifeld and Skyler J Cranmer. A theoretical and empirical comparison of the temporal exponential random graph model and the stochastic actor-oriented model. *Network Science*, 7(1):20–51, 2019.

Barbara C. Leigh, Mary R. Gillmore, and Diane M. Morrison. Comparison of diary and retrospective measures for recording alcohol consumption and sexual activity. *Journal of Clinical Epidemiology*, 51(2):119–127, 1998.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. 1987.

Dean Lusher, Johan Koskinen, and Garry L. Robins, editors. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press, Cambridge, 2012. doi: 10.1017/CBO9780511894701.

Adalbert Mayer and Steven L. Puller. The old boy (and girl) network: Social network formation on university campuses. *Journal of Public Economics*, 92(1-2):329–347, 2008.

Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pages 29–42. ACM, 2007.

Jacob L. Moreno and Helen H. Jennings. Statistics of social configurations. *Sociometry*, 1 (3/4):342–374, 1938.

Martina Morris. A log-linear modeling framework for selective mixing. *Mathematical Biosciences*, 107:349–377, 1991.

Martina Morris. Telling tails explain the discrepancy in sexual partner reports. *Nature*, 365 (6445):437, 1993a.

Martina Morris. Epidemiology and social networks:: Modeling structured diffusion. *Sociological Methods & Research*, 22(1):99–126, 1993b. doi: 10.1177/0049124193022001005.

Martina Morris and Mirjam Kretzschmar. Concurrent partnerships and the spread of hiv. *Aids*, 11(5):641–648, 1997.

Martina Morris, Ann E. Kurth, Deven T. Hamilton, James Moody, and Steve Wakefield. Concurrent partnerships and hiv prevalence disparities by race: linking science and public health practice. *American Journal of Public Health*, 99(6):1023–1031, 2009.

Brian Mustanski, Michelle Birkett, Lisa M. Kuhns, Carl A. Latkin, and Stephen Q. Muth. The role of geographic and network factors in racial disparities in HIV among young men who have sex with men: An egocentric network study. *AIDS Behavior*, 19(6):1037–1047, 2016.

Andreea Nita, Laurentiu Rozylowicz, Steluta Manolache, Cristiana Maria Ciocănea, Iulia Viorica Miu, and Viorel Dan Popescu. Collaboration networks in applied conservation projects across europe. *PLoS One*, 11(10):e0164503, 2016.

World Health Organization et al. Global incidence and prevalence of selected curable sexually transmitted infections–2008. geneva: World health organization; 2012, 2015. URL `http://www.who.int/reproductivehealth/publications/rtis/stisestimates/en/`.

Philippa E. Pattison and Garry L. Robins. Neighborhood-based models for social networks. *Sociological Methodology*, 32:301–337, 2002.

Jennifer A. Pellowski, Seth C. Kalichman, Karen A. Matthews, and Nancy Adler. A pandemic of the poor: Social disadvantage and the U.S. HIV epidemic. *American Psychologist*, (4):197–209, 2013. doi: http://dx.doi.org/10.1037/a0032694.

Anthony E. Reading. A comparison of the accuracy and reactivity of methods of monitoring male sexual behavior. *Journal of Psychopathology and Behavioral Assessment*, 5(1):11–23, 1983.

Garry L. Robins, Philippa E. Pattison, and Jodie Woolcock. Small and other worlds: Network structures from local processes. *American Journal of Sociology*, 110(4):894–936, 2005.

Garry L. Robins, Philippa E. Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2): 173–191, 2007.

Katy Robinson, Nick Fyson, Ted Cohen, Christophe Fraser, and Caroline Colijn. How the dynamics and structure of sexual contact networks shape pathogen phylogenies. *PLOS Computational Biology*, 9(6):1–15, 06 2013. doi: 10.1371/journal.pcbi.1003105.

Melanie D. Rosenberg, Jill E. Gurvey, Nancy Adler, Miranda BV Dunlop, and Jonathan M. Ellen. Concurrent sex partners and risk for sexually transmitted diseases among adolescents. *Sexually Transmitted Diseases*, 26(4):208–212, 1999.

Sharon Sassler, Katherine Michelmore, and Zhenchao Qian. Transitions from sexual relationships into cohabitation and beyond. *Demography*, 55(2):511–534, Apr 2018. ISSN 1533-7790. doi: 10.1007/s13524-018-0649-8.

Catherine Lindsey Satterwhite, Elizabeth Torrone, Elissa Meites, Eileen F Dunne, Reena Mahajan, M Cheryl Bañez Ocfemia, John Su, Fujie Xu, and Hillard Weinstock. Sexually transmitted infections among us women and men: prevalence and incidence estimates, 2008. *Sexually transmitted diseases*, 40(3):187–193, 2013.

Jacqueline Scott and Duane F. Alwin. Retrospective versus prospective measurement of life histories in longitudinal research. *Methods of Life Course Research: Qualitative and Quantitative Approaches*, pages 98–127, 1998.

Jeffrey A. Smith and James Moody. Structural effects of network sampling coverage I: Nodes missing at random. *Social Networks*, 35(4):652 – 668, 2013. ISSN 0378-8733. doi: https://doi.org/10.1016/j.socnet.2013.09.003.

Jeffrey A. Smith, James Moody, and Jonathan H. Morgan. Network sampling coverage II: The effect of non-random missing data on network measurement. *Social Networks*, 48:78 – 99, 2017. ISSN 0378-8733. doi: https://doi.org/10.1016/j.socnet.2016.04.005.

Tom A. B. Snijders. The statistical evaluation of social network dynamics. *Sociological Methodology*, 31:361–395, 2001.

Tom A. B. Snijders, Philippa E. Pattison, Garry L. Robins, and Mark S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36:99–154, 2006.

Tom AB Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40, 2002.

Arthur A. Stone, Ronald C. Kessler, and Jennifer A. Haythomthwatte. Measuring daily events and experiences: Decisions for the researcher. *Journal of Personality*, 59(3):575–607, 1991.

David Strauss and Michael Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212, 1990.

Bonnie R. Tran, Anne G. Thomas, Florin Vaida, Mooketsi Ditsela, Robert Phetogo, David Kelapile, Christina Chambers, Richard Haubrich, and Richard Shaffer. Comparisons of reported sexual behaviors from a retrospective survey versus a prospective diary in the botswana defence force. *AIDS Education and Prevention*, 25(6):495–507, 2013.

Cheng Wang, Carter T. Butts, John R. Hipp, Rupa Jose, and Cynthia M. Lakon. Multiple imputation for missing edge data: a predictive evaluation method with application to add health. *Social Networks*, 45:89–98, 2016.

Duncan J Watts. The "new" science of networks. *Annu. Rev. Sociol.*, 30:243–270, 2004.

Lance S. Weinhardt, Andrew D. Forsyth, Michael P. Carey, Beth C. Jaworski, and Lauren E. Durant. Reliability and validity of self-report measures of hiv-related sexual behavior: Progress since 1990 and recommendations for research and practice. *Archives of Sexual Behavior*, 27(2):155–180, 1998.

David P. Wilson, David G. Regan, and Jane S. Hocking. Coverage is the key for effective screening of Chlamydia trachomatis in Australia. *The Journal of Infectious Diseases*, 198 (3):349–358, 08 2008. ISSN 0022-1899. doi: 10.1086/589883.

Yoosik Youm and Edward O. Laumann. Social network effects on the transmission of sexually transmitted diseases. *Sexually Transmitted Diseases*, 29(11):689–697, 2002.

Appendix A

Proof: Onset Selection vs. Terminal Selection

In lastK designs, the average duration of ties picked by terminal selection is greater than or equal to the average duration of ties picked by onset selection: In this section we compare two variants of the lastK design: onset selection and terminal selection. We demonstrate that, for fixed K and measurement time, any tie chosen by terminal selection is either also chosen by onset selection, or else starts no later and ends no sooner (hence having longer duration) than any tie chosen by onset selection and not terminal selection.

The proof is as follows. Assume that a respondent has N ties and $N > K > 0$ (if $N \leq K$, the proof becomes trivial, because both schemes will pick all N ties). Define the sets of ties captured by onset selection and terminal selection schemes respectively as $T^O = \{t_1^O, t_2^O, ..., t_K^O\}$ and $T^T = \{t_1^T, t_2^T, ..., t_K^T\}$.

If the two schemes pick exactly the same ties, i.e. $T^O = T^T$, then any tie chosen by onset selection is also chosen by terminal selection. Thus the condition trivially holds.

Now consider the case $T^O \neq T^T$. Because $|T^O| = |T^T| = K$; there must exist ties picked by onset selection that are not picked by terminal selection, and there must be an equal number of ties picked by onset selection but not terminal selection. Formally, $\exists\ t^O \in T^O, t^T \in T^T$ that $t^O \notin T^T$ and $t^T \notin T^O$.

Define $\mathcal{O}(t)$ and $\mathcal{T}(t)$ to be the onset and terminus of a tie t. Consider a pair of ties $t^O \in T^O, t^T \in T^T$, such that $t^O \notin T^T$ and $t^T \notin T^O$. Suppose that t^T starts after t^O, $\mathcal{O}(t^T) > \mathcal{O}(t^O)$. By definition, onset selection includes the K ties the subject has most recently started. Given that t^O is chosen by onset selection, t^T which starts after t^O has to be chosen as well. This contradicts the assumption that $t^T \notin T^O$. Thus $\mathcal{O}(t^T) \leq \mathcal{O}(t^O)$.

Now suppose that t^T ends before t^O, i.e. $\mathcal{T}(t^T) < \mathcal{T}(t^O)$. Since terminal selection includes the K most recently ended ties, then the inclusion of t^T implies that t^O must be included as well. This contradicts the assumption that $t^O \notin T^T$. Thus $\mathcal{T}(t^T) \geq \mathcal{T}(t^O)$.

We have shown that, for any $t^T \in T^T, t^O \in T^O$ such that that $t^T \notin T^O$ and $t^O \notin T^T$, $\mathcal{O}(t^T) \leq \mathcal{O}(t^O)$ and $\mathcal{T}(t^T) \geq \mathcal{T}(t^O)$. Thus the durations (denoted by $\mathcal{D}(\cdot)$) of the two ties satisfy:

$$\mathcal{D}(t^T) = \mathcal{T}(t^T) - \mathcal{O}(t^T) \geq \mathcal{T}(t^O) - \mathcal{O}(t^O) = \mathcal{D}(t^O)$$

This completes the proof.

An immediate implication of this result is that the average duration of ties captured by terminal selection is greater than or equal to the average duration of ties captured by onset selection. Furthermore, the time period spanned by the ties captured by terminal selection necessarily includes the time period spanned by an equivalent onset selection design. To the extent that capturing longer-lasting ties spanning a longer period of time is a measurement objective, terminal selection is to be preferred over onset selection.

Appendix B

Derivation of the likelihood with change score

To simplify equation (4.1)

$$\mathcal{L}(\theta) = \prod_{i,j,t=1}^{N,M,T} \frac{\exp(\theta^\top S_{C_t}(y_{ijt}))}{\exp(\theta^\top S_{C_t}(y_{ijt}^+)) + \exp(\theta^\top S_{C_t}(y_{ijt}^-))}$$

we divide both the numerator and the denominator by $\exp(\theta^\top S_{C_t}(y_{ijt}^-))$.

Numerator:

because

$$\exp(\theta^\top S_{C_t}(y_{ijt})) = \begin{cases} \exp(\theta^\top S_{C_t}(y_{ijt}^+)), & y_{ijt} = 1 \\ \exp(\theta^\top S_{C_t}(y_{ijt}^-)), & y_{ijt} = 0 \end{cases}$$

then

$$
\begin{aligned}
&\frac{\exp(\theta^\top S_{C_t}(y_{ijt}))}{\exp(\theta^\top S_{C_t}(y^-_{ijt}))} \\
=&y_{ijt}\frac{\exp(\theta^\top S_{C_t}(y^+_{ijt}))}{\exp(\theta^\top S_{C_t}(y^-_{ijt}))} + (1-y_{ijt})\frac{\exp(\theta^\top S_{C_t}(y^-_{ijt}))}{\exp(\theta^\top S_{C_t}(y^-_{ijt}))} \\
=&y_{ijt}\exp\left(\theta^\top(S_{C_t}(y^+_{ijt})) - S_{C_t}(y^-_{ijt}))\right) + (1-y_{ijt}) \\
=&y_{ijt}\exp(\theta^\top\Delta_{ijt}) + (1-y_{ijt}) \\
=&\begin{cases} \exp(\theta^\top\Delta_{ijt}), & y_{ijt}=1 \\ 1, & y_{ijt}=0 \end{cases} \\
=&\exp(y_{ijt}\theta^\top\Delta_{ijt})
\end{aligned}
$$

Denominator:

$$
\frac{\exp(\theta^\top S_{C_t}(y^+_{ijt})) + \exp(\theta^\top S_{C_t}(y^-_{ijt}))}{\exp(\theta^\top S_{C_t}(y^-_{ijt}))} = \exp(\theta^\top\Delta_{ijt}) + 1
$$

Appendix C

Supplementary Figures

Fig. C.1 expands on the life event space depiction of fig. 2.1, illustrating how different events or would not be captured by different RLH designs on a sample from a larger population. The graceful degradation of lastK versus the sharp cutoff of intervalN is evident here, with the two types of designs exhibiting very different measurement profiles.

Fig. C.2 provides a detailed complement to fig. 2.9. Performance of individual samples are plotted as dots. From top to down we illustrate inferential performance of $K = 3, 4, 5, 6$; we see that inference degrades with look-back time and improves with increasing K. This plot further illustrate how inferential performance does not degrade smoothly. Within a relative short to moderate "safe" interval, all estimates stay close to the true value; outside of that interval, some samples suddenly become poor in quality. Large fraction of estimates yield infinite values when parameters associated with relatively rare events (e.g. same sex ties), a consequence of no rare events being captured within some samples.

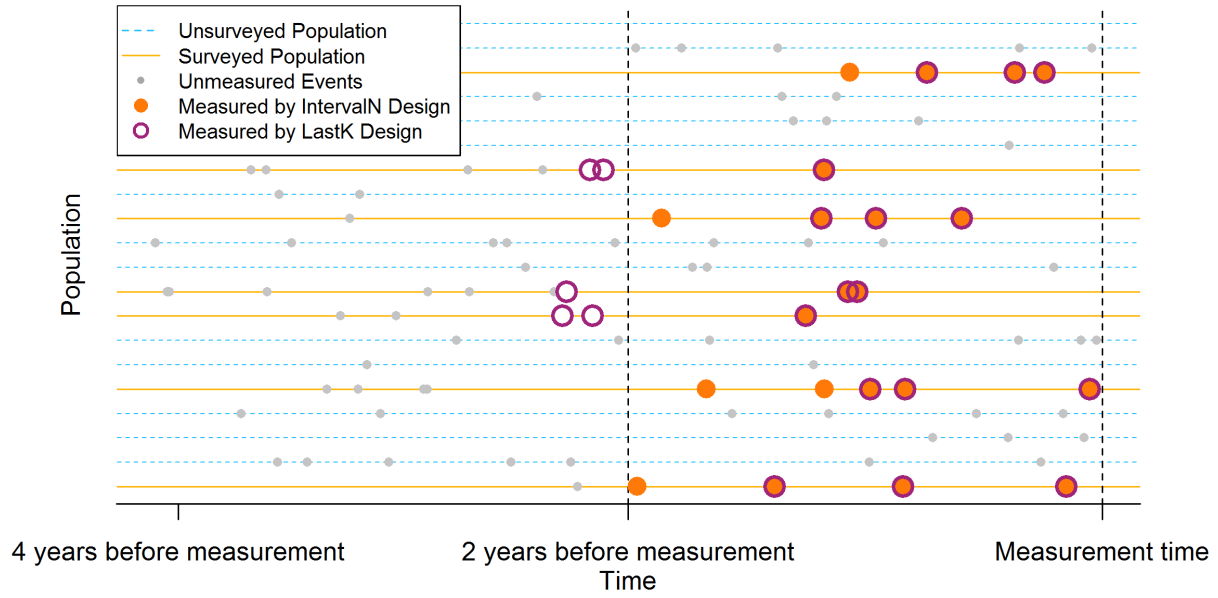

Figure C.1: An illustration of the events that are sampled by two different designs: lastK ($K = 3$) and intervalN ($N = 2$ yr.). Events closer to the measurement time are more likely to be sampled by both designs, and events that are farther away in time have less chance to be observed by either design.

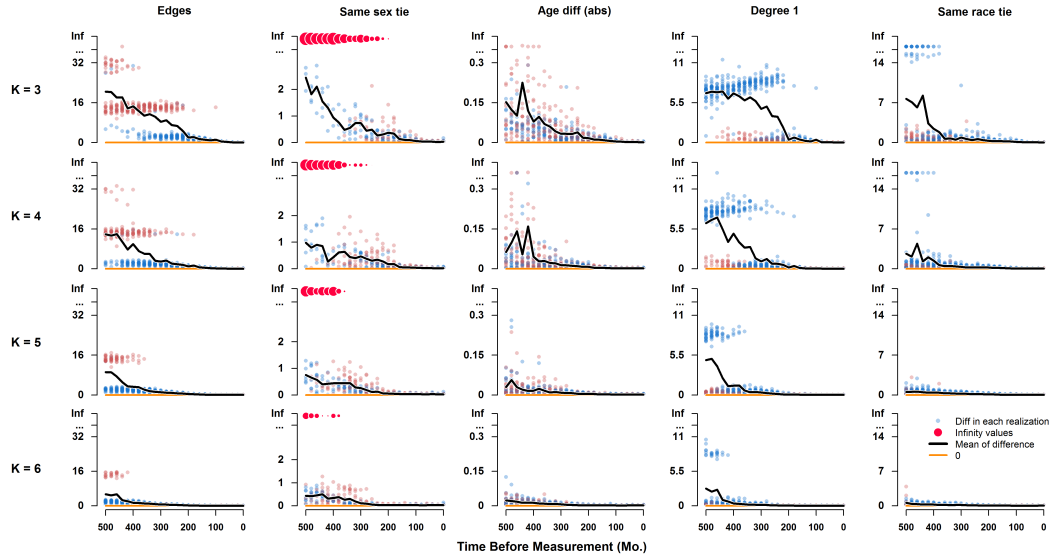

Figure C.2: Detailed results for model-based inference under the lastK designs. Vertical axis is the (absolute) error of fitted ERGM parameters for the true versus the observed networks as a function of look-back time. Points indicate individual simulation outcomes; red and blue colors are used to indicate whether the error is positive or negative respectively. The mean error is indicated by solid black lines. Circles at inf scales with the number of samples that return infinite estimates (an MLE for that parameter does not exist).